Computational analysis of primary sequence patterns in the human genome linked with regulation of gene expression and chromatin organization

Submitted By: Sameet Mehta M.Sc.

To the University of Pune in partial fulfilment of the Degree of

Doctor of Philosophy in
Zoology

Under guidance of

Dr. Sanjeev Galande
Scientist, NCCS

October, 2008

National Center for Cell Science, Pune

Contents

List of Figures

List of Tables

This thesis is dedicated to my family.

Acknowledgments

I thank Dr. G. C. Mishra, Director, National Center for Cell Science, for making available a good working environment and infrastructure for my work. I thank Dr. Sanjeev Galande, my guide, for accepting me as a student, guiding me and helping me in my difficult times. I would like to thank from the bottom of my heart all the members of the SG Lab at the National Center for Cell Science, Pune, for their constant support. I would also like to thank Dr. Hemant Purohit and his lab members at the National Environmental Engineering Research Institute, Nagpur. I thank the Council for Scientific and Industrial Research, which provided me with fellowship during the tenure of my thesis. I also thank Dr. Dhanajay Raje a former member of Dr. Purohit's lab at NEERI. I thank Dr. D. G. Kanhere, Director, Center for Modeling and Simulation, University of Pune, for allowing me to work at the CMS, and providing me with financial assistance. I thank Dr. Mihir Arjunwadkar, CMS, for his constant encouragement and support during the very crucial finishing stages of the thesis. I thank Mr. Abhay Parvate, for stimulating discussions and fresh perspective on nearly all the problems that I could discuss with him. I thank all the administrative staff at the NCCS, NEERI, and CMS for making my working environment highly conducive to work. I thank my brother Ranjan, who has been a source of constant inspiration and a driving force. I thank my wife Shruti, who stood by me during my highs and lows. Last but not the least I thank my mother who was a constant emotional support.

DECLARATION

I hereby declare that the work for the thesis entitled "Computational analysis of primary sequence patterns in the human genome linked with regulation of gene expression and chromatin organization" submitted for Ph.D. degree to the University of Pune, has been carried out at the National Center for Cell Science, University of Pune under the supervision of Dr. Sanjeev Galande. The work is original and has not been submitted in part or full for any degree or diploma to this or any other university.

Dr. Sanjeev Galande
(Supervisor)

Sameet Mehta
(Ph. D Candidate)

CERTIFICATE

It is hereby certified that the work incorporated in the thesis entitled, "Computational analysis of primary sequence patterns in the human genome linked with regulation of gene expression and chromatin organization" submitted for the degree of *Doctor of Philosophy* by **Sameet Mehta** has been carried out under my supervision. Materials obtained from other sources have been duly acknowledged in the thesis.

Date:

Dr. Sanjeev Galande,
(Supervisor)

Dr. G. C. Mishra,
Director,
National Center for Cell Science,

Chapter 1

Introduction: Literature Review and Background

1.1 Background

Revolution in the DNA sequencing technology over the past decade led to phenomenal increase in its throughput and cost reduction. However, the burgeoning sequencing data also led to unprecedented set of problems for biologists. During the same period there were significant advances in the electronics that led to increased power of computing. Today we are at a stage where a single genome may be sequenced in a matter of few hours to few days, a task which used to require months to years not so long ago. Additionally advances in mathematics and statistics yielded very powerful analytical tools and techniques to deal with the large amount of sequence data (for both DNA and proteins) generated as a result of high throughput sequencing techniques. The important question is then to understand what all these sequences mean? Initial studies involving sequence analyses were focused on defining a measure of 'similarity' within sequence(s) to determine their phylogeny. Moreover, current studies use these and similar sequence alignment techniques to identify 'common' or homologous sequences between various inter-/intra- genomic regions. Multiple techniques for sequence alignment were introduced in early 90s (1). These have been improved over the years with addition from techniques honed and perfected in diverse fields of science (2, 3, 4). Techniques from diverse disciplines such as the Language theory are also being applied to biological sequence analysis problems (5). Furthermore, inside a cell the functional state of the genome is in form of an orderly nucleoprotein complex called the chromatin (6). The genomic DNA and DNA-bound histones form a major component of this chromatin, in addition to chromatin remodeling

complexes and various non-histone DNA binding proteins. The hierarchical packaging of chromatin poses an interesting question regarding the features/information in the primary genomic sequence that act as a driving force for the compact assembly of chromatin.

In addition to the genomic sequence itself, plethora of information about the transcriptome[1], and the proteome[2] was also generated. This led to new disciplines of studies that is collectively known as *omics*.

In this chapter we discuss the established and current computational methodologies for analysis of the genomic sequences and their functional elements. We briefly discuss a few areas of contemporary interest, and various computational techniques used in these studies.

1.2 Challenges

With more and more genomes being sequenced, one of the biggest challenges was and continues even today, to be annotation of the genome(s). The genomic DNA sequence is basically (nearly) an endless string(s) of letters A, C, G, and T. Although it is widely accepted that genomic sequence features regulate gene expression, delineating these features is a very complicated problem. As a further complication to the problem, most of the higher eukaryotic genomes are made up of a number of repetitive elements which makes it very difficult to interpret results based on statistical analysis of DNA sequences. These repeats can lead to number of false positive and false negative results.

Multiple Sequence Alignment The first question is usually identification of the organism of origin for the DNA under study. This question is especially pertinent when the origin of the DNA is not a well defined source. To this end, traditionally a multiple sequence alignment approach is used. One of the earliest approaches for comparison of two sequences was the FASTA algorithm (1). These techniques are used most commonly in the study of flora (bacterial population) of unknown environments. The most common method to answer the question of organism of origin of the DNA under investigation is to align the obtained (test) sequence with known sequences. Depending on the distance of the unknown sequence from sequences with known origin a fair estimate can be made about the origin of the unknown

[1]Complete RNA complement of a given genome
[2]Complete protein complement of a given genome

sequence. Such methods usually involve some sort of multiple sequence alignment. Multiple sequence alignment usually means global alignment wherein the algorithm tries to match the entire length of the query sequence (the unknown sequence) with entire length of the target sequence (the known sequence(s)). This is achieved by adding gaps to either of the sequence(s) as required. However, over the years better algorithms have been designed to increase the speed and efficiency of global sequence comparisons. One of the most well known of these algorithms is the `CLUSTAL` (2). Over number of years the sequence comparison algorithms have evolved, such that today we have a wide choice for programs/algorithms for multiple sequence alignment. Following programs are well known and widely used, `MAFFT` (7), `T Coffee` (3), `MUSCLE` (8), `DIALIGN-T` (4), etc. Most of the improvements have been in efficiency. Moreover, with widespread availability of the World Wide Web (Internet), increasing number of sequence analysis platforms are available online.

Motif Detection Over a number of years, motif detection has attracted a lot of attention. Motif detection is carried out in nucleic acid (DNA and RNA) as well as the amino acid (Protein) sequences. In both cases, the aim of motif detection is to arrive at a continuous conserved sub-sequence amongst the given set of sequences. Most of the motif detection algorithms were originally written for protein sequences with which they yield more reliable results due to a long alphabet (20 amino acids). However, the algorithms have now been adapted to use nucleic acid sequences as inputs.

The motif detection algorithms can be classified as either enumerative or probabilistic. The enumerative method is a thorough method wherein all possible combinations of a sequence of given length are generated, and score is assigned to each combination. Each such occurrence is counted and most occurring sub-sequence is considered to be a motif. One of well known example of enumerative algorithm for motif detection is `weeder`, and its web implementation (9).

In probabilistic method(s), each position in a motif is considered. At the start of analysis, each position is assigned equal probability of having any amino acid/base. With each subsequent occurrence of that n-mer, these probabilities are re-calculated and reassigned so that the score for a particular base (in DNA and RNA)/acid (amino- acids of proteins) varies according to the probability of its occurrence. The end result is generation of a position

specific probability matrix (PSPM) (see page 85 for further explanation). One of the well known and widely used probabilistic algorithm is the `MEME`, (10).

Using enumerative motif finding algorithms becomes computationally untenable as the length of the motif increases. This is especially true if the motif detection is being carried out for protein sequences. The possibilities in the nucleic acid sequences increase with a factor of 4^n where n is the length of the motif. Moreover for proteins the possible combinations increase with factor of 20^n. For detecting a motif of length 8 – 10 the possible number of combinations already approach figures that are computationally untenable even for high-end super computers. In such cases probabilistic motif detection algorithms are preferred.

The probabilistic methods can generate a high number of false positive and false negative results. To circumvent this problem it is usually advised that multiple algorithms be used on a given data set and also that same algorithm be used multiple times (11, 12). It also depends heavily on the statistical tests employed to determine whether a detected motif is really statistically significant or not. This particular problem is further exacerbated because the genomes are replete with various repetitive elements such as the LINES, SINES and Alu-repeats. Additionally, genome is a multi-hierarchical complex structure with numerous long-range and short-range correlations that affect interpretation of the results (13). A few tools are available today to detect such long-range correlations and to emulate them in generated sequences such that a valid question about the significance of motifs may be posed (14). For an excellent comparison of the various methodologies and tools available for motif detection, refer to Tompa *et al.* (15). Similarly, Hu *et al.* have discussed limitations and potentials of the available motif discovery algorithms (16) and also proposed an ensemble algorithm (17) for motif detection.

Probabilistic Motif Detection As mentioned earlier, in probabilistic motif detection we usually generate a PWM (Position Weight Matrix). In a particular case of transcription factor binding site detection the PWM is generated using data from known binding sites for a given transcription factor. The effectiveness of this approach depends on the models ability to accurately formalize the regularities found in the confirmed sites (18, 19). However, this approach relies on two strong assumptions, viz., a) all the positions within a site are

independent, and b) all the binding sites of a transcription factor are variations of the same sequence, these assumptions are not easily satisfied.

Whole Systems Biology (Systems Biology) As the volume of biological information is increasing, so is the awareness of looking at the information from multiple perspectives. Investigations focused at understanding the entire system rather than what happens to a single gene/protein are undertaken by many laboratories. This approach is broadly known as "Systems Biology". For an excellent disposition on the definition of systems biology see the commentary by Kirschner (20). Traditional approach of a biologist usually involves reduction of the system to its components. Thus, one usually selects a gene of interest and digs around that gene, until total effect of that gene on the organism/system under study is elucidated. In the systems biology approach, whole organism/pathway/biological phenomenon is studied as a single system, with internal connections. The basic premise of systems biology is *functional value of a system is greater than the sum of its parts.* These approaches are especially relevant in todays post-genomic era[3]. Moreover, it is now generally accepted that viewing 'data' (genome sequence and its features and annotations) from multiple perspectives gives better insights and understanding of the biological processes that they affect.

These approaches are especially necessary and important in data analyses of the microarray (high throughput gene expression profiling studies), ChIP-on-chip (CoC) (genome-wide binding studies) etc. One of the most important aim of such studies is reconstruction of the Transcription Regulatory Networks (TRN)s.

Briefly, in microarray experiments, cDNA from experimental and test samples are labeled with different dyes, and hybridized to slides with DNA fragments corresponding to known, predicted and hypothesized genes. Depending on the intensities of the different dyes, an inference can be made about overexpression/repression of several thousand genes at a time. Analyzing such data can show internal-correlations and/or anti-correlations in the expression patterns of multiple genes for a given treatment (test). In the CoC experiments, the chromatin is cross-linked inside the cells under control and test conditions. A chromatin iimmunoprecipitationn (ChIP) is performed using antibodies against a protein of interest.

[3]Generally the time when the cost and time for genome sequencing have reduced significantly and it is well within reach for many laboratories to in principle have multiple genomes sequenced.

After reversing the cross-links in these different samples, the DNA is labeled with different colors, and hybridized to slides spotted with representative DNA from various areas of the target genome. The readout directly represents the occupancy of a given region of DNA by protein of interest under different conditions. Both gene expression profiling experiments and CoC experiments are popularly known as "high throughput" analyses.

As the whole system is studied the complexity of these problems is immense. So the workers have tried to break the problems into manageable small systems. Lee *et al.* have studied the protein-protein interactions in the *Saccharomyces cereviseae*, and demonstrated the emergence of the TRNs through the studies of these interactions (21).

Recent Advances With huge advances in the computer sciences, technological innovations in the fields of electronics and computer building, better and powerful computers are available for lesser cost. Simultaneously, more and more researchers from across the disciplines have been able to contribute analysis techniques honed and matured in their respective branches of science to analyze biological data. People have applied well known techniques such as Simulated Annealing, Support Vector Machines, Genetic Algorithms, Neural Networks, for predicting of patterns in the biological data (22, 23, 24). More specifically Aerts *et al.* have used these techniques to decipher cis-regulatory motifs (25, 26). Recently various techniques from the Language Theory have also been used to study, detect, and analyze patterns in biological data (27, 28, 5).

Furthermore, *meta-data*[4] analysis is also gaining root in high throughput experiments such as the gene expression microarray, ChIP-on-chip etc. A few well known commercial software programs available to perform analyses of such high throughput data are `Bibliosphere`[5] and `LitInspector` available at the `GenoMatix`[6]. Few open source programs and application interfaces to query the meta-data databases, like `BioConductor` a software suite written in and for the `R`[7] open-source statistical environment. Similarly there are modules in `BioPerl`[8] that allow querying of `GO` databases[9].

[4]Meta-data is the data which is available over and above the sequence itself. It may include genomic location, presence of genes/transcripts, putative regulatory sequences etc.

[5]`Bibliosphere` is now available for free use but is not open sourced

[6]`http://genomatix.org`

[7]`http://cran.r-project.org`

[8]`http://bioperl.org`

[9]GO (Gene Ontology) database is a functional annotation database for the known and predicted genes.

1.3 Case Studies

We illustrate the use of the sequence analysis through following case studies.

1.3.1 Cis regulatory motif/module detection

The problem of coordinated regulation of gene expression has intrigued biologists for many decades (29). Promoters have a known biological function[10] and this property makes these class of sequence motifs very interesting. Moreover, the promoters themselves also have an inherent ability to regulate expression of genes (30). One of the foremost large scale studies was carried to determine putative promoter elements in the human genome (16). The Eukaryotic Promoter Database (`EPD`) is an annotated non-redundant collection of eukaryotic PolII promoters, experimentally defined by a transcription start site (TSS). Access to promoter sequences is provided by pointers to positions in the corresponding genomes. Promoter evidence comes from conventional TSS mapping experiments for individual genes, or, from mass genome annotation projects (31).

Various approaches have been taken to study these special sequences (32, 33). It should be noted however, that there are no well-known properties of the regulatory regions, like the coding regions. Further, these regulatory regions are not distributed in the genome in uniform fashion, neither are they distributed randomly (34). There is no regular spatial distribution. The consensus regulatory elements are degenerate more often than not, which makes delineation of the exact consensus difficult or impossible.

In higher eukaryotes like mammals, the regulatory regions can be divided into two general categories. The *proximal* regulatory regions e.g. promoters are typically near the $5'$ end of a gene. The *distal* cis-regulatory modules (CRM)s include the enhancers. The CRMs may be located far upstream, within or downstream of a gene. These are difficult to identify because they do not necessarily have any specific location with respect to the transcription start site.

Following are the biological phenomena (35) which necessitate use of high performance computing to address the problems of detection of TRNs and understanding biology, viz.,

[10]The promoters usually act as the primary docking site for the PolII. However, there is still no consensus about the positioning of the promoter for all the genes. Usually -1 to -1000 from the TSS (Transcription Start Site) is considered to be the basal promoter

- multiple transcription factors (TF)s tend to regulate gene activity in distinct regulatory modules
- individual TFs usually have multiple binding sites within a regulatory module
- binding sites within a regulatory module tend to be spatially clustered

Moreover, mere presence of Transcription Factor Binding Site (TFBS) at a particular upstream region does not necessarily mean that the TF is actually bound there. Hu *et al.* have clearly shown that there is a difference between binding of a TF to the target site and effect of the bound TF on expression (36). All the studies of the regulatory code of the genome should therefore be seen in this light (37, 38).

In general the methodology for the the determination of conserved cis-regulatory modules involves analysis of the genomic DNA. The sequences under study are usually selected based on their affinity for the transcription factor of choice. In not so distant past, EMSA (Electrophoretic Mobility Shift Assay) used to be technique of choice for determining the target sites of the DNA binding proteins. ChIP (Chromatin immunoprecipitation) allows high throughput screening of the regions that bind specifically to the protein of interest. Additionally, techniques such as SELEX (systematic evolution of ligands by exponential enrichment) (39), that afford a high throughput and high affinity assay for selection of the protein-binding DNA sequences. SELEX in particular is now preferred to arrive at specific protein binding DNA sequences on account of its relative ease of use and ability to screen very large number of putative protein binding sequences. However, SELEX suffers from a disadvantage; the sequences obtained from these are good binders, but it is difficult to compare their relative affinities.

It should be noted however, that statistical computational recognition of regulatory regions is desirable but very difficult. Following factors contribute to complexity of the problem, rendering it nearly untenable, viz.,

- Lack of known properties, like open reading frames, non-uniform codon usage in coding sequences.
- Degeneracy of the TFBS, and small length of the consensus binding site, making it difficult to accurately detect these sites.

- Complicated and non-regular structure of the regulatory regions. There are no consistent sequence motifs in the regulatory regions. These regions comprise a collection of diverse TFBS.

One of the latest and model studies to determine the cis-regulatory modules has been by Hu *et al.* (36). The overall scheme followed in the study was defined earlier by Lee *et al.* (21). For the first time a panel of knock-out yeast strains wherein each strain lacked a specific transcription factor were used together for the expression microarray analysis as well as ChIP-on-chip studies. The study shows that there is more to regulation of the gene-expression by TFs than mere binding and clearly discriminates between these two phenomena, and in the process enables building of better models for the re-construction of the transcriptional regulatory networks.

Some of the computational approaches to solve this problem will be discussed here. The most common methods employed to this end are use of known TFBS, use of information from the DNA sequence itself (content based methods) and more recently the phylogenetic foot-printing.

Use of Known TFBS One of the most well known sources for TFBS is the `TRANSFAC` database (40). The `TRANSFAC` database contains known consensus binding sites for the TFs, position-weight matrix (PWM)[11]. The PWM is basically representation of the TFBS, giving probability of each nucleotide at each position. This representation takes care of the 'degeneracy of the consensus TFBS. Looking for presence of such TFBS is one of the simplest approaches to arrive at a rudimentary transcription regulatory network (TRN). However, this simplistic model of one transcription factor to one gene does not depict the immense complexity of the TRNs, inside a living cell/organism. This simplistic model is enhanced by addition of constraints and/or parameters such that the model closely resembles the real system. It should be noted here that on account of the complexity of the biological systems, and the TRNs inside a living cell, many times it is impossible to replicate the entire complexity of a living system (cell) in the models. Once such parameter added is that of the distance between the transcription factor binding sites (41). Such a model may be further tuned to

[11]For discussion on Position Specific Probability Matrix (PSPM) a variant of PWM see Chapter 3

take into account occurrence of pairs of TFs or more TFs together, i.e. *composite elements* (42). Another method `PEAKS` that works from first principles is proposed by Bellora *et al.* (43). This method utilizes the information about the genomic landmarks like the Transcription Start Site (TSS), and the genomic sequences around such a landmark to show presence of specific TFBS. This approach however suffers from the fact that the CRMs are not necessarily clustered around the genomic landmarks. On the contrary there is ample literature available wherein the CRMs have been seen to be present far upstream or far downstream of the TSS.

Content Based These methods rely on detection of the differences in the base-composition of the regulatory and the non-regulatory regions. These are the most popular methods for discovery of new motifs. Ohler *et al.* (44) describes use of Interpolated Markov Chains for promoter detection. The same group has also used physical properties of the DNA sequences and information derived from the sequence itself to predict promoter sequences (45). Local word (8 mer) frequencies were used to determine CRM in *Drosophila melanogaster* developmental genes (46). Narlikar *et al.* have shown that combining known information about the sequence with application motif finding algorithms improves the efficiency and accuracy of the motif finding algorithms (47). Orlov *et al.* have used statistical measures of structure of genomic sequences: entropy, complexity, and position information for elucidation of promoter sequences and other CRMs from genomic sequences (48). Thus it may be seen that it is possible to use information from the DNA sequence itself to arrive at CRMs.

Phylogenetic Footprinting Another well known approach is based on recognition of the regulatory DNA based on evolutionary conservation. This is known as phylogenetic footprinting (49, 50). Phylogenetic footprinting is an approach to find functionally important sequences in the genome that relies on detecting their high degrees of conservation across species (51). Mutations are more likely to be disruptive if they appear in functional sites, resulting in a measurable difference in rates of evolution between functional and non-functional genomic segments (52). It should be noted that for biological function of the TF inside a cell it depends on many additional factors than presence of mere binding site(s). Lenhard *et al.* (52) describe a software tool to identify conserved regulatory elements by comparative

genome analysis. This approach has been successfully utilized by Allenede *et al.* (53), to detect enhancers across multiple species. An algorithm that uses ChIP-on-chip data and the phylogenetic foot-printing has been discussed which can use data sets from various species to arrive at a regulatory sequence (54). Nimawegen described the mathematical and statistical framework for integrating various motif/module finding algorithms and discussed the phylogenetic footprinting methods (55). It may be reasoned that use of multiple algorithms for elucidation of regulatory motifs/modules will provide a better confidence on the deduced motifs (56, 57). As shown by Shi *et al.* (58), when multiple methodologies are combined with data from multiple species, more confidence can be allotted to the detected motifs. Moreover, with addition of the GO (Gene Ontology) (59) search terms some insight can be obtained about the role these motifs play in regulation of transcription. This is one of the methods that allows de novo identification of regulatory regions from genomic sequences.

This method depends heavily on alignment-like algorithms and hence its accuracy depends on the evolutionary distance of the species being compared. Further, this approach offers little information about the specific functions of the conserved sequences. Moreover, there is no consensus opinion for the number of sequences required to reliably extract regulatory region(s).

Summary It should be noted that the genomic sequence carries signals needed for function over and above the regulation of gene expression such as the large-scale chromatin remodeling and all these signals are essentially superimposed on one another. Furthermore, in all these methods especially the probabilistic methods for determination of the CRMs there is always a necessity for presence of control sequences. Usually these control sequences are obtained by in silico generation of the DNA sequences based on certain rules. These rules are usually derived from the sequences under study. The basic assumption for generating these sequences is that the DNA sequence is identical and independent distribution (IID) of the nucleotides. This assumption may not always hold true. For a detailed discussion about this topic of generating proper control datasets refer to chapter 3.

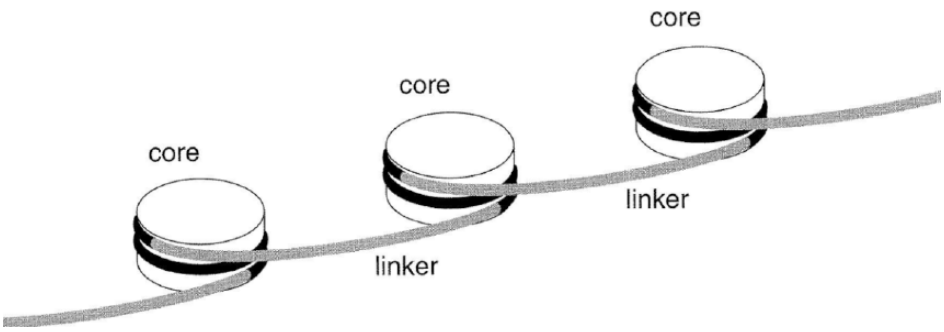


Figure 1.1: Cartoon of the DNA wound around nucleosome particle.

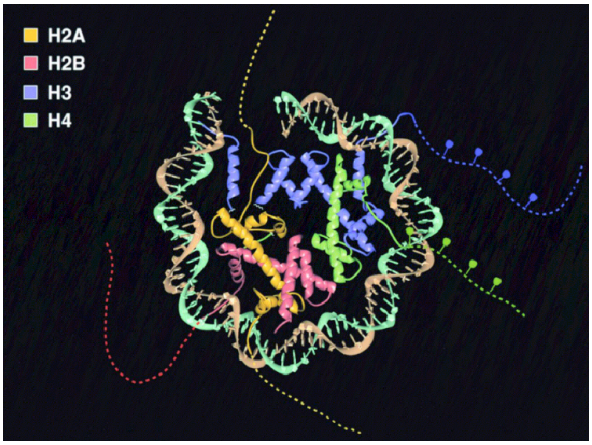


Figure 1.2: Nucleosome Crystal Structure (adopted from Luger *et al.* (61)).

1.3.2 Nucleosome Positioning

What is a Nucleosome? A nucleosome is the basic subunit of the chromatin. The nucleosome is fundamental to DNA coiling and gene regulation. It serves as a basis for chromatin condensation. The primary event in gene activation may be the modification of histones and the resulting decondensation of large chromosomal domains (60). Nucleosomal DNA in *Saccharomyces cereviseae* is 165 bp long, of which 146 bp wrap around the histone octamer in 1.65 turns. The histone octamer is composed of two copies of each histone H2A, H2B, H3, and H4 has been highly conserved throughout evolution (61). Genomic DNA sequences show considerable variability in their binding affinity to the histone octamer, and this variability contributes to determining the location and distribution of nucleosomes (62, 63).

Wrapping of DNA into a nucleosome influences transcription factor binding to its cognate sites, and thus the positions of nucleosomes in eukaryotic genomes contribute to gene

regulation (64). The genomic DNA wrapped around the nucleosomes gives the "beads on a string" appearance to the chromatin. In Figure 1.1, the puck shaped white cylinders are the histone cores. The black string around the histone cores in the nucleosomal DNA which is in contact with the histone cores, and the grey DNA is the linker DNA.

The high resolution crystal structure for the nucleosome core particle was solved in 1997 by Luger *et al.* (61). The histone tails in Figure 1.2 are denoted by dashed lines of appropriate length and the lollipops denote the known modifications of the tails. The crystal structure of the nucleosome core particle revealed the details of the histone-DNA interaction(s) (61). These interactions are confined to the phosphodiester backbones of the DNA strands. A set of contacts is made every 10 base pairs where the minor groove on the double helix faces inwards. Electrostatic interactions and hydrogen bonding with the DNA phosphates as well as nonpolar contacts with the deoxyribose groups are observed (61, 60).

The tails of the histones extend well outside the core-complex itself and are important for the functioning of the nucleosome in transcription regulation. From the literature, the nucleosome may be described as follows,

- A nucleosome may be defined as a histone octamer made up of two copies each of H2A, H2B, H3, and H4, with DNA wound on the outside.
- Each histone is organized into two domains: a central fold, which lies within and constrains the DNA superhelix, contributing to the compact core of the nucleosome; and an unstructured amino-terminal tail, which extends outside the core and provides a basis for interaction and regulation.
- A chain of nucleosomes is coiled in a chromatin fiber through interactions of the histone tails with adjacent nucleosomes and additional proteins; these interactions may be modulated by post-translational covalent modification(s) of the tails.
- Chromatin-remodeling complexes clear nucleosomes from enhancers, promoters, and other specific protein-binding sites in chromatin.
- Many DNA-binding regulatory proteins repress transcription by recruiting histone acetyl-transferases or deacetylases, which act on nearby nucleosomes.
- Stable repression of transcription by the formation of heterochromatin is based on the

> nucleosome. Interaction of histone tails with silencing proteins starts at special sites and spreads along the chromosome to form repressive structure that may persist through many cell generations.

Thastrom *et al.* (65) have studied and highlighted the nucleosomal locations on known DNA sequence with high affinity for formation of nucleosome. Their studies revealed that of the 147 bases that are wrapped around the histone core, the central 71 bases are most important. These make contact with the H3-H4 dimer whereas the flanking 38 bases make contact with the H2 histones. The free energies of all these interactions play role in the stability of the nucleosome and its positioning. Recently it has been shown that there is stretching and kinking of the DNA as it wraps around the nucleosome core particle (66). Recently high-throughput techniques like ChIP-seq have also been used to determine architecture of nucleosome in terms of histone variants (67).

For in-depth discussion of structure of the nucleosomes please refer to Kornberg and Lorch (60) and other reviews (68, 69, 70, 71).

Role of Nucleosomes There is ample evidence to show that nucleosomes inhibit transcription. It is therefore believed that the nucleosomes limit access to the DNA, whereas the access to the intervening linker DNA is more relaxed. In 1989 Csordas proposed an interesting hypothesis as follows (72),

> *Introns were used in course of evolution for the organization of eukaryotic genes within repeating units of nucleosomes, since the distinct DNA conformations of the nucleosome core particle and of the linker region, respectively, represent a constraint for the positioning of genes.*

It has also been shown that nucleosomes can act as activators or repressors to the gene in vicinity, depending on the which cognate sites are available for binding to the transcription factors. White *et al.* (73) have shown that in the yeast minor sequence variations lead to dramatic changes in the way in which nucleosomes pack against each other. This has important implications for our understanding of the formation of higher order chromatin structure and its modulation by post-translational modifications. Yuan *et al.* have shown that at least in yeast functional transcription factor binding sites were found in the linker

region (74). Moreover, they also found that depletion of nucleosomes in the transcription start sites is seen in many yeast promoters (74). Pryciak & Varmus (75) have shown that sequences that favor nucleosomes also favor integration of HIV PIC. In the same study it was shown that the chromatin organization at PolII promoters consists of a nucleosome-free region approximately 200 base pairs upstream of the start codon flanked on both sides by positioned nucleosomes. Further, nucleosome-free sequences were evolutionarily conserved and were enriched in poly-deoxyadenosine or poly-deoxythymidine sequences and most occupied transcription factor binding motifs were devoid of nucleosomes, strongly suggesting that nucleosome positioning is a global determinant of transcription factor access (74). Davey *et al.* (76) have studied the nucleosome positioning signals in the mouse and human H19 imprinting control region(s) (ICR). Senkinger *et al.* (77), have shown that nucleosomes occupy DNA and mask fortuitous TF binding sites in the genome. It has been shown recently that the nucleosomes regulate chromatin compaction thus play important role in repression of gene expression (78). Similarly, Gutirrez *et al.* have shown that during activation of chromatin the chromatin remodeling factors SWI/SNF evict nucleosomes and enable the TFs to access the DNA (79).

It thus appears that nucleosome(s) and any changes to the DNA sequences targeted preferentially by the nucleosome(s), have a profound effect on the regulation of transcription.

Sequence Determinants of Nucleosome Positioning Physical basis for nucleosome friendly DNA sequences has been known (80). Nucleosome positioning for the most part takes advantage of the intrinsic structural mechanics of the double helix. Moreover, the DNA duplex has a tendency to bend towards the minor or the major groove (*roll*) much more easily than in a direction along the longer base-pair axis (*tilt*). In the nucleosomal DNA there is even greater preference for *roll* over *tilt*. The *roll* in nucleosomal DNA contributes to smooth bending into either groove and to kinking into the minor groove (81, 82). Kinking is almost never seen in other protein-DNA complexes (83). The kinking in the minor groove is almost exclusively seen at the CA•TG border steps (83). Furthermore, it is known that CA•TG steps have a preference for low-*roll*/high-twist in a YCAR[12] context (84), and most nucleosomal CA•TG steps with negative *roll* values are preceded by a pyrimidine and

[12]Y represents a pyrimidine and R a purine

followed by a purine. A strong anti-correlation between *twist*and *roll* has been observed earlier (85, 86). Sequence-specific constraints of the sugar phosphate backbone contribute primarily to the conformational variability of protein-bound DNA. In particular the *roll* angles in CA•TG steps cover a span of more than 40 degrees (from -21 degrees to +23 degrees). This remarkable variability allows CA•TG to act as 'hinge' and allow them to be on the "inside" or "outside" of a nucleosome core. During the protein-DNA interaction, especially with TFs, most of the free energy contribution comes from specific interactions between protein side-chains and DNA bases. For such bindings the protein has to accommodate structural properties of only a limited number of nucleotides, and YR dimers have been selected over course of evolution as most frequent sequence elements to "fit" DNA around the protein because of their unique conformational properties (87, 83). When a TF binds to DNA in sequence specific manner, the interaction is very localized, as the recognition site is short. This is achieved efficiently within a small portion of conformational space available to the positive *roll* angles possible with YR dimers (87, 83). In contrast, the nucleosome core has to wrap a longer piece of DNA regardless of its sequence. This is a "global" optimization problem and a large conformational variability has to be explored in the DNA to achieve minimum free energy for the binding. Part of the solution is similar utilization of high *roll* angles of the YR dimers. However, nucleosomal DNA has certain characteristics that are only partially employed in B-DNA oligomers and specific protein-DNA complexes: (1) Higher overall flexibility of all dinucleotides; (2) extremely tight coupling of twistand *roll*angles; (3) negative *roll* angles in CA•TG steps. Stein & Bina have shown with elaborate experiments that VWG[13] sequences with 10 base periodicity serve as powerful nucleosome positioning signals (88).

It has been shown that nucleosomes are not distributed statistically uniformly or stochastically on DNA but rather are organized in specific arrangements that have been implicated in mechanisms controlling gene expression (89, 90). For reading more on crystallographic studies of oligonucleotide sequences please read (91, 92). Bendable DNA is favorable sequence whereas stiff DNA is unfavorable sequence for nucleosome formation. Ioshikhes *et al.* (93) have shown that there are indeed specific base preferences as specific positions in the

[13]V means any nucleotide other than T, W means A or T

nucleosomal DNA. Pazin *et al.* have shown that nucleosomes are dynamic and mobile rather than static and that a DNA binding factor is continuously required for the maintenance of nucleosome positioning (94). The mechanism of nucleosome positioning and its functional consequences have been discussed extensively (95, 96).

Specifically, the AA and TT dinucleotide as specific positions enhance binding of the DNA sequence to the nucleosome core. Lowry & Widom have elucidated sequence rules of natural nucleosomal DNA with a strong statistical significance (97). It is known that poly-A regions are stiff and are not good targets for nucleosome formation. On the other hand, sequences containing AT dinucleotides are easy to bend and sequences containing AT dinucleotides with 10-base periodicity have high affinities for the histone octamers. Thastrom *et al.* (98) have shown that different dinucleotides have different free energies for binding to nucleosome core. Their analysis shows special significance for nucleosome positioning of a motif consisting of approximately 10 bp periodic placement of TA dinucleotide steps. They further show that contributions to histone binding and nucleosome formation from periodic TA steps are more significant than those from other periodic steps such as AA (=TT), CC (=GG) are more important than those from the other YR steps (CA (=TG) and CG), which are reported to have greater conformational flexibility in protein-DNA complexes even than TA (98).

Yuan *et al.* (74) have identified genomic nucleosome positioning sites in the entire chromosome III of the yeast *Saccharomyces cereviseae*. Fernandez & Anderson (99) have shown that local DNA structures are important for positioning and that single base-pair changes at nucleosome formation sites could have profound effects on those genomic functions that depend on ordered nucleosomes. Moreover, it has also been shown that specific dinucleotides are important for nucleosome positioning. Similarly, recently Segal *et al.* have proposed a genomic code for nucleosome positioning (100).

Protein Determinants of Nucleosome Positioning It should be noted that in addition to the genomic sequences some proteins also pay an important role in the assembly of nucleosomes. Nucleosomes are positioned during the replication-coupled (RC) nucleosome assembly by the chromatin assembly factor 1 (CAF1) chaperone complex, which is tethered to the replication processivity clamp (PCNA) (101). Further, the positioning sequences only contribute to the probability that a site will be occupied by the nucleosomes, the actual maintenance of

the nucleosome at a position is however dictated by the action of DNA-binding proteins and nucleosome-remodeling complexes (64). It is known that RC coupled nucleosome assembly is brought about by CAF1, and is sequence independent (102). Nucleosome positions and stability are also affected by presence of the H1 linker histone (103, 104). Studies have also assigned anti-silencing function protein 1 (Asf1) an important role in nucleosome assembly. It forms a complex with H3–H4 (105). A proposed role of Asf1 in nucleosome is discussed in detail by Henikoff (106). In yeast it has been demonstrated that the Asf1 and acetylation of H3K56 has a direct role in RC coupled nucleosome assembly and dis-assembly (107).

The nucleosome structure and function also changes according to the nature of histones incorporated. Most of the core histones have variants and these variants are usually associated with specific functions in the cells. Furthermore, different histone chaperones and nucleosome assembly pathways are associated with different histone variants (106). For the effect of variant H3.3 on epigenetic inheritance of active chromatin see page 19. For a discussion of epigenetic modifications of the nucleosome-histones please see page 20.

Computational Determination of Nucleosome Formation Potential As mentioned earlier, nucleosome positioning plays a very important role in the transcription regulation. Moreover, there are definite sequence signature(s) in the genome, which affect positioning of the nucleosome and hence affect regulation of transcription, directly or indirectly. Thus these sequences assume a regulatory role in addition to the promoters, enhancers, silencers, and other *cis*-regulatory sequences. As such it has been of considerable interest to be able to harness the computing power to predict such nucleosome positioning sequences. Such attempts have been scant. Most interesting studies in this direction have been by Levitsky *et al.* (108, 109). They used various algorithms to arrive at nucleosome formation potential. Moreover, they also studied the differences between the nucleosome formation potential of the house keeping genes vis-a-vis the tissue specifically expressed genes and demonstrated that promoters governing the expression of these to groups of genes are indeed different in terms of the nucleosome formation potential (108). Peckham *et al.* (110) used Support Vector Machines for prediction of the genomic sequences with high nucleosome formation potential. They used a training dataset from Segal *et al.* (100) with nucleosome promoting and inhibiting sequences. The algorithm arrived at the periodicity that is known to be

present in the nucleosome forming sequences. They successfully demonstrated that genomes encode an intrinsic nucleosome organization and that this intrinsic organization can explain approximately 50% of the *in vivo* nucleosome positions. Furthermore, they also concluded that this nucleosome positioning code may facilitate specific chromatin functions including transcription factor binding, transcription initiation and even remodeling of the nucleosomes themselves (110).

Summary It is thus clear from the wealth of literature available that nucleosome is a fundamental unit of chromatin organization. It has a unique structure and affinity to bind with DNA. Moreover, depending on the sequence and physical properties like bendability of the DNA the affinity of nucleosome core for binding with DNA changes. In particular TA steps[14] are important nucleotides that favor formation and maintenance of the nucleosomes. Rules have been elucidated from the sequences that promote and inhibit nucleosome formation, especially in the yeast genome. The presence of the nucleosomes affect chromatin structure, access of the transcriptional machinery to the DNA and hence the regulation of transcription itself. With advent of computers there have been attempts to use modern algorithms like the support vector machines to predict nucleosome formation potential of the DNA sequences. Further studies in this area are necessary to arrive at a comprehensive set of rules which will aid in analysis of the genomic DNA sequence analysis problem, and possibly give an insight into regulation of transcription.

1.3.3 The Epigenetic Code

Over number of years, it was noted that the functional state of chromatin is heritable. Such epigenetic information in the form of histone modifications, is characterized by complexity, diversity and an overall tendency to respond to changes in genomic function rather than to predict them (111). Epigenetic research is about understanding the heritable regulation of gene expression that is not directly coded in the genomic DNA (112). We know that replicating cells are able to maintain their identity from generation to generation. This essentially requires that the pattern of gene expression that defines a cell type is maintained identically through the generations. This is referred to as *cellular memory* (113, 114, 115).

[14] a step usually means a 10 base-pair periodicity

When the epigenetic inheritance occurs between generations of cells it is known as *mitotic inheritance*. Whereas when the epigenetic inheritance occurs between generations of the species it is known as *meiotic inheritance*. Both are seen in the eukaryotic world. Epigenetic information can sometimes be inherited across multiple generations (116). There are three biochemical mechanisms that are commonly referred to as epigenetic (112),

- DNA methylation,
- histone modifications, and
- binding of non-histone proteins such as the Polycomb and trithorax group complexes

The question then is, will this information be passed on in form of histone modifications? This is a crucial question, because if true, then changes to the histones that are induced by metabolic or environmental influences on the modifying enzymes involved will change not only the cells initially subjected to these influences but also their progeny (111). Nightingale *et al.* (111) have proposed that the terms histone code and epigenetic code be distinguished as follows viz.,

Histone code To refer to the combinations of modification(s) that are known (at least in principle) to be involved in ongoing cellular processes

Epigenetic Code To refer to the putative heritable code that might be responsible for the cellular memory.

One of the first demonstrations of the non-genomic heritability of the state of chromatin was the *position-effect variegation* (PEV) as seen in *Drosophila melanogaster* (106). It has been shown that the heterochromatin is dynamic in small-time-scales and inaccessible to the transcription factors (106, 117). Recently it has been shown that the micro-RNAs also effect epigenetics in mammals, though the mechanism(s) involved are ill understood (118). In this section we review the literature pertaining to epigenetic code and its basis in the genomic sequences in a cell.

Histone Modifications The histones of the nucleosome core are subjected to a wide, and ever increasing variety of post-translational modifications (119). A standard nomenclature has been suggested to describe these modifications (120). Many histone modifications have

been documented viz., methylation (121, 122), acetylation (123), phosphorylation (124, 125), ubiquitination (126), ADP-ribosylation (127), sumoylation (128), deamination, and proline isomerization (129). Of all the known histone modifying enzymes the kinases and the methyl transferases are most specific. Histone modifications play a functional role because specific proteins bind to modified histones, moreover, some modifications are required for binding of these proteins (130). The resultant effect on the transcription may be a function of the binding of such factor, or more likely binding of multiprotein-small-RNA complexes to the modified histone tails. There are known correlations between the covalent modifications and combinations-of-modifications of the core histones and their functional implications. Di- and tri- methylation of H3K9, together with de-acetylation of the histones, is hallmark of transcriptionally silenced chromatin (131, 132). Similarly, a combination of hyper-acetylation and H3K4me3 is a characteristic of transcriptionally active chromatin. Barski *et al.* studied and mapped histone methylation patterns in the entire genome using the *Solexa*-technology (133). They have used high-throughput methodology and elucidated typical methylation patterns at various genomic landmarks such as enhancers, promoters, silencers, boundary elements, and transcribed regions. Similarly it has been shown that hyperacetylation of H3K9, H3K14, and H4 is positively correlated with H3K4 methylation, and all these changes mark a transcriptionally active state for chromatin (134, 135). Recently Mikkelsen *et al.* (136) mapped chromatin states in lineage-committed cells and established a nearly 1:1 correlation between various histone modifications and transcriptability. Few studies also demonstrate that the successive modifications of the histones can control the high order chromatin structure and hence regulation of transcription (137, 138, 139). Further, the effects may vary depending on whether the modification is in a single nucleosome core or is distributed over neighboring nucleosome cores. The histone modifications themselves can affect the chromatin in many ways from transcription activation, gene silencing, DNA repair, and cell-cycle progression (140). Thus, although the functional effects of the modifications might be identical, the distribution of the modifications along the tails may vary greatly. In one case the modifications may be present on a single nucleosome core tail. Whereas in the other, the modifications may be distributed over a chromatin territory. It is known that the distribution of these histone modifications across the domain is non-trivial.

The acetylated isoforms of the histones are seen predominantly in the promoters rather than the body of the genes. The differentially methylated isoforms of H3K4 differ in their distribution. H3K4me3 is consistently most enriched at the beginning of the genes, H3K4me2 in the middle and H3K4me1 at the end (111). H3K36me3, and H3K79me3 are also consistently enriched across coding regions (111).

The chromatin modifications act either by disrupting contacts of the modified nucleosome cores with some factors, or by recruitment of some proteins/factors to the nucleosome, and in effect changing the higher order chromatin structure (119).

With high-throughput methods like the ChIP-on-chip, it is possible to map the histone modifications on a genome-wide scale. The ChIP-on-chip analyses have shown that the modification sites are spread across entire genome (141, 142, 143). Moreover, Schübeler *et al.* and Bernstein *et al.* showed a positive correlation between various histone modifications is conserved across the species (141, 144). However the ChIP-on-chip approach has its own limitations. It can detect the modification status over a range of nucleosomes or even on a single nucleosome, but it cannot determine the modification status of different histones within the same nucleosome. The only way to find this is Mass Spectrography, however, the requirement of digestion in this technique limits its potential. However, a new 'top-down' approach of first identifying the protein and then digesting it that may allow studying modifications on intact proteins (145, 119).

Role of Histone Variants The stability of the nucleosome is affected by incorporation of the histone variants (146). The incorporation of the variant histones marks chromatin into distinct regions (146). This is exemplified best by difference between the centromeric and non-centromeric chromatin, which differ only in presence or absence of the histone H3 variant, CenH3. In the non-centromeric regions the H3 variant present universally is H3.3. It is believed that this counterpart of the CenH3 is a central player in maintaining epigenetic inheritance (147, 148). It has been shown that deposition of H3.3 occurs primarily in transcriptionally active chromatin and gene regulatory sites (149, 150, 151). Only 4 amino acids distinguish H3 from H3.3, of these 3 are in the core which prevent H3 from being deposited during replication (106). These core residues are exposed in the soluble Asf1-H3-H4 escort complex, which presents H3-H4 and H3.3-H4 to the CAF1 and Hir1 (152), also see page 17.

There are major differences in the modification patterns of H3 and H3.3. Curiously, removal of the N-terminal tail of the H3.3 has no effect on its incorporation indicating that the tail modifications are not important for H3.3 incorporation (153, 149). Active genes are enriched in both H3.3 and modifications that mark 'active chromatin such as di- and tri- methylation of H3K4 (H3K4me2, H3K4me3) etc., and is depleted in markers of inactive chromatin such as H3K9me2. The silent chromatin is depleted in both (154, 155, 156). In yeast rapid turn over of the histones was observed in the promoters of actively transcribed genes (157). This rapid turnover is very high in the promoters and progressively diminishes over the downstream region of the gene (158, 159, 107). A similar gradient is observed in the occupancy of the H3.3 in *Drosophila* promoters (149). Even though the yeast lacks H3.3, its H3 is classified as H3.3, and uses both the CAF1 and Hir1 pathways for nucleosome assembly, implying that the pathways of nucleosome assembly that dictate "transcribed" versus "repressed" chromatin states are conserved (106). Recently Moorman *et al.* (160) have demonstrated that the "hotspots" for transcription factor binding sites correspond broadly with H3.3 presence, suggesting that nucleosome turnover is a general mechanism for the transcriptional machinery to gain access to the chromosomal targets.

In addition to the histone H3 variants mentioned here other variants are also seen the eukaryotic world. H2A has known variants namely H2A.Z, MacroH2A, H2A-Bbd and H2A.X (161). MacroH2A is enriched on the human inactive X chromosome and is enriched specifically in regions that undergo X-inactivation (162). H2A-Bbd is a histone H2A variant specifically depleted in the inactivated X-chromosome. This phenomenon demonstrates how the variants of a single histone play role on epigenetic inheritance, where one variant marks active chromatin and the other inactive chromatin (163). Additionally, the H4 and H1 histone variants play an important role in tight packaging of the sperm DNA (161).

DNA Modifications In addition to the histones and their variants the DNA itself is subject to modifications. Especially the CpG islands are known targets of methylation, and this modification is known to play important role in the regulation of transcription (164, 165). Further it is known that modified DNA is more susceptible to alterations during replication (166). In the same line it has been shown that the fidelity of replication of modified DNA is very low (10^{-3} mutations per base pair) (166). The mutation rates at

modified bases are much higher than expected mutation rates of 10^{-8} per base for unmodified bases during normal cell division (167). Non-CpG methylation has an established functional role in plants (168, 169). It has also been observed in mammals in early development, and embryonic stem cells, but it is significantly decreased in differentiated cells (170). It has been shown earlier that epigenetic changes in the genome occur during development, along with changes in the expression patterns of the transcription factors (171, 172, 173). Further, *imprinting*, which is a kind of epigenetic inheritance, plays a very important role in development (174, 175). There are specific disorders known which are linked with disruption of either maternal or paternal imprinting e.g. Prader-Willi Syndrome, Angelman Syndrome, and Beckwith-Wiedemann syndrome (176, 177). The abnormality in the epigenetic marks are now believed to be reason for many human afflictions (178, 179).

There is a close link between DNA CpG methylation and histone modification, with hypermethylated regions being rich in histone marks that define a transcriptionally repressed chromatin (180, 111). Whether the H3K methylation at Lysines other than at position 4 drives DNA methylation or its is the DNA methylation that drives formation of transcriptionally repressed chromatin is not properly understood (111). Moreover, there is evidence to suggest that this mechanism may vary from system to system. Studies in Arabidopsis have shown that H3K9 methylation drives DNA methylation (181). However, other studies have shown that there are indeed loci with intermediate properties of heterochromatin in which transcription downregulation is inherited in a manner similar to constitutive heterochromatin, although the loci are associated with opposing histone marks H3K4me2 and H3K9me2 (182, 183).

Role of Non-histone Proteins Non-histone proteins also affect chromatin structure. The *polycomb* group (PcG) of proteins and the *trithorax* group[15] (trxG) of proteins act as proteins that either bring about the epigenetic modification of the chromatin, or they read the epigenetic modification of the chromatin and help in effecting the consequences of such modifications (112). Both the PcG and the trxG act throughout the genome (106).

It has been shown that the PcG actually brings about methylation of DNA by Viré *et al.*

[15]These were identified as mutants of the same name in *Drosophila melanogaster*. Subsequently homologues were found in higher animals.

(184). The PcG acts though *Enhancer-of-zeste* subunit of the *Polycomb repressive Complex 2* (PRC2), and methylates the histone H3 at lysine 27 (185). This H3K27-methylation keeps the chromatin in repressed state (186). By default PcG maintains repressive state of chromatin. The trxG proteins help reverse this and prevent silencing of the genes during early development (187). The trxG includes motif-specific DNA binding proteins (e.g., Zeste), nucleosome remodellers (e.g., Bramha and Kismet), H3K4 methyltransferase (e.g., Trithorax), and a H3K4 demethylase (e.g., Lid) (188, 189). This diversity of the trxG function is also seen in the vertebrates, and analogous proteins with similar functions have been identified (190, 191). It is also possible that there are as yet un-discovered trxGs, which are intimately involved with the epigenetic reading and writing mechanism that maintain active chromatin (106). It has been shown that the histone H3–H4 chaperone (Asf1) is encoded by suppressor of PEV and interacts with Brahma *in vivo* and *in vitro* (192). Although the PcGs and the trxGs function in nearly diametrically opposite fashion, they recognize and bind to same sites called the PREs or PRETREs (Polycomb response element trithorax response element) (106). The binding of the PcGs or trxGs occurs irrespective of the transcriptional state of the target homeotic promoter (193). Another protein involved with epigenetic processes is RbAp48 (also known as MSI1), which is a component of various nucleosome-assembly complexes as well as PRC2, the NURF nucleosome remodeller and other chromatin associated complexes (194). The exact role is sketchy, however it is speculated that RbAp48 containing complexes act on partially disassembled nucleosomes during dynamic processes such as replication and transcription (106).

Heritability of Chromatin State The problem of assigning a heritable role to histone modification(s) is difficult. For an excellent discussion of mechanisms of epigenetic inheritance read (195). In case of DNA methylation the problem is well understood. Replication of methylated DNA leads to hemi-methylated sites on daughter DNA which are the preferential substrate of DNMT1, a DNA methyltransferase, associated with replication machinery (196). The explanation of the heritable modifications of the histones forming the nucleosome cores is not that straightforward or well understood (195), especially since the heritable modifications and the non-heritable modifications co-exist. The techniques such as ChIP-on-chip or ChIP are gross techniques wherein the average state of the cells is represented. The only way

around this limitation is to assay levels of histone modifications in a preferably synchronized culture through mitosis. A study in this direction has been published by Valls *et al.* (197). Moreover, it has been pointed out that a specific role for a single modification does not constitute a code (198). This is true because typically the experiment is setup to investigate a particular modification of histone (111). This also remains to be a limitation of ChIP-on-chip experiments and their analysis because they focus on ongoing transcription and are not set up to detect a "heritable" code. The replication of "chromatin" involves DNA synthesis and nucleosome assembly, which occur coordinately in the cell (195). The replication coupled (RC) nucleosome assembly is the process in which nucleosomes are formed on newly synthesized daughter strands. Examination of the nucleosome core structure revealed that H3 and H4 tetramer could theoretically be split into H3–H4 dimer (195). In context of the RC chromatin assembly, it might be possible that the nucleosome disassembles to give two H3–H4 dimers. Each of this dimer may go to a daughter strand, and form hybrid nucleosomes with new core histones associated with RC nucleosome assembly (199). Alternatively, nucleosomes may be divided intact between the two daughter DNA molecules (199, 200). The semiconservative model is attractive in a sense since it can explain some of the observed phenomena. The conjecture is that the epigenetic information obtained within each old H3–H4 dimer could be used as a template to copy information onto the new H3–H4 dimer, thus fully replicating original nucleosome and also the chromatin status with modification (195).

It has been proposed that the nucleosomes split along H3–H3 dimerization interface during replication resulting in formation of half nucleosomes, which act as template for the addition of new half nucleosome (201). Later studies have showed that the $(H3–H4)_2$ tetramers were inherited intact (202, 163). Furthermore, there is no known mechanism that can replicate a modification between two half nucleosomes (203). There is some evidence to accept this semi-conservative replication in a small subset of regulatory sites that transmit epigenetic memory (204). However, after extensive studies, there is still confusion about whether the nucleosomes disassemble and are divided (202). This phenomenon is also explained by another theory which uses the rapid turnover of the nucleosomes at active promoters. It is proposed that process of histone turnover at regulatory elements perpetuates itself, thus maintaining chromatin in a constitutively active state (163). Another possibility is described by Henikoff

(106), which I quote here,

> *The silent state would be default. CAF1, together with Asf1, would assemble H3 nucleosomes at replication that are enriched in silent modifications and deficient in active modifications. Conversely, replication-independent incorporation of H3.3 by other chaperones, such as HirA, and disassembly by Asf1 would occur at sites of transcriptionally active chromatin and regulatory elements, process that is set in motion by action of transcription factors. Over the course of cell cycle, actively modified H3.3 would accumulate at active genes and regulatory elements, and the random partitioning to daughter chromatids would favor perpetuation of the active state. Efficient CAF1-dependent assembly behind the replication fork would help to perpetuate the silent state over broad regions, whereas the local turnover process that results in histone replacement would cause H3K27 methylation to be lost, thus counteracting silent chromatin.*

It is believed that the rapid turnover of the nucleosomes is carried out by ATP-dependent nucleosome remodellers (205, 150). It is also possible that during remodeling some nucleosomes are occasionally evicted, transiently exposing DNA and allowing PcGs and/or DNA binding proteins to find their target sites (106). Then the continued local presence of nucleosome remodellers would result in another cycle of remodeling, nucleosome depletion and histone replacement at the said sites. This model has been proposed to explain the short occupancy times of the transcription factors on the target DNA *in vivo* (206, 207). This would lead to reduced nucleosome density and DNAseI hypersensitivity, especially if replacement of nucleosomes is a slow process (150). Indeed it has been observed that the PREs are deficient in nucleosomes as compared to their flanking regions (193, 186). Mito *et al.* have shown that this relationship is true for the entire genome (150).

Computational Epigenetics The "Computational" epigenetics comprises of two broad approaches to study where the computing power is employed viz., i) Data analysis, ii) Prediction.

Data Analysis To understand the computational aspect in the epigenetics studies, it is essential to familiarize oneself with the techniques used. As mentioned earlier, some of

the popular approaches for studying epigenetics are ChIP-on-chip, ChIP-seq, and bisulphite sequencing (112). For a detailed discussion about the factors involved in design and analysis of ChIP-on-chip experiments, see Buck and Leib (208). In a modified format of ChIP-on-chip, called the methyl-DNA Immunoprecipitation (MeDIP), the antibodies are used against an epigenetic modification of the DNA (209). The ChIP-seq is a variant of ChIP-on-chip. It uses high-throughput DNA sequencing rather than tiling arrays for detecting differences between sample and control DNA (136, 133). The ChIP-seq has advantages over ChIP-on-chip, i) Data normalization is not an issue because the results obtained are absolute read counts. ii) With advances in sequencing technology, the experiments are very cost effective. The frequently used bisulphite sequencing method exploits the ability of bisulphite to convert DNA methylation state of cytosines into a methylation-dependent SNP (210, 211).

All the techniques used to study epigenome data generate vast amount of data which require efficient ways of data processing and quality control. In analysis of a ChIP-on-chip data the challenge is to derive a ranked list of over-represented genomic regions from raw intensities (212, for detailed discussion see). Usually a three step process is used as follows (213),

1. microarrays are quantile-normalized and standardized to a common median intensity,
2. a Wilcoxon rank sum test is applied locally on a sliding window to test for differential hybridization and to derive a Z-score for each probe.
3. significant probes are merged into regions of overrepresentation if sufficiently close to each other, and these regions are ranked by their combined Z-score.

Various approaches from Hidden Markov Models, linear models, probabilistic models were developed to improve the spatial resolution of the peaks (214, 215, 216). Several toolkits are now available in the public domain (not necessarily open sourced) to deal with ChIP-chip datasets, we list a few here for posterity e.g. `TileMap`, `ChIPOTle`, `Ringo`[16] (217, 218, 219). This problem of peak-detection is still unsolved. A framework has been proposed by Du *et al.* (220) to identify most informative regions.

The problems of analysis with ChIP-seq are slightly better understood than those of ChIP-on-chip. The key step in this analysis is fast and accurate mapping of the short sequence to the

[16] a `BioConductor` package

reference genome. This is usually done using `blastn`[17] (1), or `BLAT` (221). Unlike the probe intensities in the ChIP-on-chip, the ChIP-seq corresponds to a genomic fragment bound to the target. So normalization is virtually not required. However, analysis of this type of data comes with its own set of pitfalls. The process of mapping tags to the reference genome can bias the analysis toward genomic regions with unique and complex sequence patterns. This is because short sequencing reads that may partially overlap with low-complexity regions or with other interspersed repeats stand a higher chance of being discarded for lack of unique genomic alignment (112).

We discuss the Data analysis part only briefly because it is beyond the scope of this thesis.

Prediction In addition to the analysis of the ChIP-on-chip and the ChIP-seq data, some basic sequence analysis at the genomic sequence level is also part of the analysis of the epigenetics/epigenome. Furthermore, it is always useful to build statistical model(s) of the epigenetic information available, and improve these models as more and more information is available (112). The work in this direction was started by predicting promoters (222). Moreover, for a detailed discussion on prediction of cis-regulatory modules/motifs please see section 1.3.1, on page 7. Though the efforts in promoter prediction in highly annotated genomes e.g. Human genome have slowed down, meta-analysis of known promoters is now on the rise. Smith *et al.* (223) have looked at presence of the tissue-specific regulatory elements in the promoters. Similarly, Bulcke *et al.* and others have looked into re-constructing the TRNs by mining multiple genomic information resources (224).

Another area of interest specifically related to epigenetics/computational epigenetics is prediction of the CpG islands. Many CpGs actually overlap with promoter regions (225). CpGs play a general role as regulator of chromatin remodeling. It is also known that the CpGs are most common targets for methylation. Therefore prediction and identification of bona-fide CpG islands is important. The criteria currently used for such assessment were proposed by Bock *et al.* (226). The prediction of methylation (regions where CpG methylation is highly probable) is conceptually easier because the methylation patterns are largely independent of tissue as compared to other epigenetic marks (112). It has been observed by several workers that DNA methylation prediction for sequences derived from different tissues such

[17] Available at `http://www.ncbi.nlm.nih.gov/BLAST`

as lymphocytes (227), and brain (228) have comparable results. Furthermore, the sequences that are predicted to be CpG islands are also predominantly the targets for DNMT1 (229, 230, 112).

In addition to the use of computers and machine learning techniques to predict and/or estimate CpGs and their methylation, attempts have also been made at predicting post-translational modifications of the histones computationally. Li *et al.* have demonstrated a software tool that successfully predicts the histone acetylation sites[18] (231).

The nucleosome positioning and its prediction is yet another area related to the prediction of the epigenome from the primary sequence. For a discussion on the topic of nucleosome positioning see section 1.3.2. By virtue of the definition of epigenetics (see page 19), it eminent that nucleosome positioning would pay in important role in regulation of transcription. Analysis of the primary sequences to predict nucleosome positioning and effects of the sequences on DNA and histone modifications is gradually gaining importance.

Summary In summary, epigenetics is important for understanding variety of biological phenomena. Both histone modifications and DNA modifications are intimately linked and affect each other (232). With new and high-throughput techniques such as the ChIP-on-chip, ChIP-seq, etc. arises immediate need for development of analytical, mathematical, and statistical tools to deal with this burgeoning data. Furthermore, with availability of highly annotated genomes, the task will be to put all the information together in a context to arrive at proper inference from such experiments. All single experiments in case of epigenetic analysis are important because each experiment contributes towards the overall knowledge. The problems in epigenetics are being understood only now, and with availability of more data the picture will hopefully become clear. It is quite apparent from the literature that the primary genomic sequences play a role in modulating and controlling the epigenetics of the genome. Though it is not very clear if the effects of the sequences are direct (sequence-specific) or indirect (more related with charge distributions being detected and recognized rather than the sequence itself). The precise role of the primary genomic sequences in directing epigenetic modifications is poorly understood and needs to be investigated further.

[18] available at `http://bioinformatics.lcd-ustc.org/pail/`

1.4 Closing Remarks

In the past few pages we have discussed various aspects of the genome that need to be studied to understand regulation of transcription (see pages 11 and 19). The genomic sequences play a very important role in regulation of all nuclear processes and "molecular biology" of the nucleus. For a long time, the role of the genomic sequences *per se* was not well understood. However, with advances in technology and availability of sequences for many genomes, our understanding of the crucial role of the sequences is improving. Previously, regions of genomic DNA which could not be assigned a canonical function were usually classified as "junk" DNA. Our understanding today tells us that there is probably no "junk" DNA in the genome. The genome actually functions as a multi-hierarchical dynamic nucleoprotein-complex called 'chromatin'. Chromatin includes genomic DNA and histones as major components in addition to non-histone DNA binding proteins and chromatin remodeling complexes. The hierarchical packaging of chromatin poses an interesting question regarding the features/information in the primary genomic sequence that act as a driving force for the compact assembly of chromatin. In this thesis, I have attempted to search for motifs in DNA that play important role(s) in selected biological processes that affect chromatin remodeling and hence regulation of gene expression. The thesis will also discuss some important considerations that should ideally be taken into account when designing and performing bioinformatics analysis of genomic DNA sequences. The probable role of the DNA sequences to direct tissue-specific expression of genes is also explored with few approaches that hold promise.

References

[1] W. R. Pearson. Using the FASTA program to search protein and DNA sequence databases. *Methods Mol Biol*, 24:307–31, 1994.

[2] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80, 1994.

[3] C. Notredame, D. G. Higgins, and J. Heringa. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1):205–17, 2000.

[4] A. R. Subramanian, J. Weyer-Menkhoff, M. Kaufmann, and B. Morgenstern. DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*, 6:66, 2005.

[5] D.B. Searls. The language of genes. *Nature*, 420(6912):211–217, 2002.

[6] N. Gilbert and B. Ramsahoye. The relationship between chromatin structure and transcriptional activity in mammalian genomes. *Brief Funct Genomic Proteomic*, 4(2):129–42, 2005.

[7] K. Katoh, K. Kuma, H. Toh, and T. Miyata. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, 33(2):511–8, 2005.

[8] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–7, 2004.

[9] G. Pavesi, F. Zambelli, and G. Pesole. WeederH: an algorithm for finding conserved regulatory motifs and regions in homologous sequences. *BMC Bioinformatics*, 8:46, 2007.

[10] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*, 34(Web Server issue):W369–73, 2006.

[11] A. Chakravarty, J. M. Carlson, R. S. Khetani, and R. H. Gross. A novel ensemble learning method for de novo computational identification of DNA binding sites. *BMC Bioinformatics*, 8:249, 2007.

[12] J. M. Carlson, A. Chakravarty, C. E. DeZiel, and R. H. Gross. SCOPE: a web server for practical de novo motif discovery. *Nucleic Acids Res*, 35(Web Server issue):W259–64, 2007.

[13] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. Simons, and H. E. Stanley. Statistical properties of DNA sequences. *Physica A*, 221:180–92, 1995.

[14] P. W. Messer and P. F. Arndt. CorGen–measuring and generating long-range correlations for DNA sequence analysis. *Nucleic Acids Res*, 34(Web Server issue):W692–5, 2006.

[15] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–44, 2005.

[16] J. Hu, B. Li, and D. Kihara. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res*, 33(15):4899–913, 2005.

[17] J. Hu, Y. D. Yang, and D. Kihara. EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences. *BMC Bioinformatics*, 7:342, 2006.

[18] B. Georgi and A. Schliep. Context-specific independence mixture modeling for positional weight matrices. *Bioinformatics*, 22(14):e166–73, 2006.

[19] J. H. McVey. Staden Plus. *Methods Mol Biol*, 25:171–9, 1994.

[20] M. W. Kirschner. The meaning of systems biology. *Cell*, 121(4):503–4, 2005.

[21] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.

[22] T. K. Rasmussen and T. Krink. Improved Hidden Markov Model training for multiple sequence alignment by a particle swarm optimization-evolutionary algorithm hybrid. *Biosystems*, 72(1-2):5–17, 2003.

[23] D. Barker. LVB: parsimony and simulated annealing in the search for phylogenetic trees. *Bioinformatics*, 20(2):274–5, 2004.

[24] M. Perez-Enciso. Multiple association analysis via simulated annealing (MASSA). *Bioinformatics*, 22(5):573–80, 2006.

[25] S. Aerts, G. Thijs, B. Coessens, M. Staes, Y. Moreau, and B. De Moor. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*, 31(6):1753–64, 2003.

[26] S. Aerts, P. Van Loo, G. Thijs, H. Mayer, R. de Martin, Y. Moreau, and B. De Moor. TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. *Nucleic Acids Res*, 33(Web Server issue):W393–6, 2005.

[27] Y. Sakakibara. Grammatical inference in bioinformatics. *IEEE Trans Pattern Anal Mach Intell*, 27(7):1051–62, 2005.

[28] T. Head. Formal language theory and DNA: An analysis of the generative capacity of specific recombinant behaviors. *Bulletin of Mathematical Biology*, 49(6):737–759, 1987.

[29] J. H. Kelly and G. J. Darlington. Hybrid genes: molecular approaches to tissue-specific gene regulation. *Annu Rev Genet*, 19:273–96, 1985.

[30] T. Waleev, D. Shtokalo, T. Konovalova, N. Voss, E. Cheremushkin, P. Stegmaier, O. Kel-Margoulis, E. Wingender, and A. Kel. Composite Module Analyst: identification of transcription factor binding site combinations using genetic algorithm. *Nucleic Acids Res*, 34(Web Server issue):W541–5, 2006.

[31] C. D. Schmid, R. Perier, V. Praz, and P. Bucher. EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res*, 34(Database issue):D82–5, 2006.

[32] L. Marino-Ramirez, J. L. Spouge, G. C. Kanga, and D. Landsman. Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res*, 32(3):949–58, 2004.

[33] S. Sonnenburg, A. Zien, and G. Ratsch. ARTS: accurate recognition of transcription starts in human. *Bioinformatics*, 22(14):e472–80, 2006.

[34] I. Abnizova and W. R. Gilks. Studying statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the eukaryotic genomes. *Brief Bioinform*, 7(1):48–54, 2006. Excellent introduction and indepth commentary on the subject of the statistical properities of regulatory DNA sequences.

[35] M. I. Arnone and E. H. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–64, 1997.

[36] Z. Hu, P. J. Killion, and V. R. Iyer. Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet*, 39(5):683–7, 2007.

[37] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

[38] L. O. Barrera and B. Ren. The transcriptional regulatory code of eukaryotic cells–insights from genome-wide analysis of chromatin organization and transcription factor binding. *Curr Opin Cell Biol*, 18(3):291–8, 2006.

[39] Y. Choo and A. Klug. Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc Natl Acad Sci U S A*, 91(23):11168–72, 1994.

[40] E. Wingender, P. Dietze, H. Karas, and R. Knuppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res*, 24(1):238–41, 1996.

[41] V. J. Makeev, A. P. Lifanov, A. G. Nazina, and D. A. Papatsenko. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res*, 31(20):6016–26, 2003.

[42] E. Wingender, A. E. Kel, O. V. Kel, H. Karas, T. Heinemeyer, P. Dietze, R. Knuppel, A. G. Romaschenko, and N. A. Kolchanov. TRANSFAC, TRRD and COMPEL: towards a federated database system on transcriptional regulation. *Nucleic Acids Res*, 25(1):265–8, 1997.

[43] N. Bellora, D. Farre, and M. Mar Alba. PEAKS: identification of regulatory motifs by their position in DNA sequences. *Bioinformatics*, 23(2):243–4, 2007.

[44] U. Ohler, S. Harbeck, H. Niemann, E. Noth, and M. G. Reese. Interpolated markov chains for eukaryotic promoter recognition. *Bioinformatics*, 15(5):362–9, 1999.

[45] U. Ohler, H. Niemann, Liao Gc, and G. M. Rubin. Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinformatics*, 17 Suppl 1:S199–206, 2001.

[46] A. G. Nazina and D. A. Papatsenko. Statistical extraction of Drosophila cis-regulatory modules using exhaustive assessment of local word frequency. *BMC Bioinformatics*, 4:65, 2003.

[47] L. Narlikar, R. Gordan, U. Ohler, and A. J. Hartemink. Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics*, 22(14):e384–92, 2006.

[48] Y. L. Orlov, R. Te Boekhorst, and I. I. Abnizova. Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. *J Bioinform Comput Biol*, 4(2):523–36, 2006.

[49] D. L. Gumucio, H. Heilstedt-Williamson, T. A. Gray, S. A. Tarle, D. A. Shelton, D. A. Tagle, J. L. Slightom, M. Goodman, and F. S. Collins. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Mol Cell Biol*, 12(11):4919–29, 1992.

[50] L. Duret and P. Bucher. Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol*, 7(3):399–406, 1997.

[51] Z. Zhang and M. Gerstein. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J Biol*, 2(2):11, 2003.

[52] B. Lenhard, A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg, and W. W. Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2(2):13, 2003.

[53] M. L. Allende, M. Manzanares, J. J. Tena, C. G. Feijoo, and J. L. Gomez-Skarmeta. Cracking the genome's second code: enhancer detection by combined phylogenetic footprinting and transgenic fish and frog embryos. *Methods*, 39(3):212–9, 2006.

[54] M. T. Friberg. Prediction of transcription factor binding sites using ChIP-chip and phylogenetic footprinting data. *J Bioinform Comput Biol*, 5(1):105–16, 2007.

[55] E. van Nimwegen. Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics*, 8 Suppl 6:S4, 2007.

[56] T. Vavouri, K. Walter, W. R. Gilks, B. Lehner, and G. Elgar. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol*, 8(2):R15, 2007.

[57] G. K. McEwen, A. Woolfe, D. Goode, T. Vavouri, H. Callaway, and G. Elgar. Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. *Genome Res*, 16(4):451–65, 2006.

[58] W. Shi, W. Zhou, and D. Xu. Identifying cis-regulatory elements by statistical analysis and phylogenetic footprinting and analyzing their coexistence and related gene ontology. *Physiol Genomics*, 31(3):374–84, 2007.

[59] Creating the gene ontology resource: design and implementation. *Genome Res*, 11(8):1425–33, 2001.

[60] R. D. Kornberg and Y. Lorch. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, 98(3):285–94, 1999.

[61] K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*, 389(6648):251–60, 1997.

[62] H. R. Drew and A. A. Travers. DNA bending and its relation to nucleosome positioning. *J Mol Biol*, 186(4):773–90, 1985.

[63] S. C. Satchwell, H. R. Drew, and A. A. Travers. Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol*, 191(4):659–75, 1986.

[64] O. J. Rando and K. Ahmad. Rules and regulation in the primary structure of chromatin. *Curr Opin Cell Biol*, 19(3):250–6, 2007.

[65] A. Thastrom, L. M. Bingham, and J. Widom. Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. *J Mol Biol*, 338(4):695–709, 2004.

[66] M. S. Ong, T. J. Richmond, and C. A. Davey. DNA stretching and extreme kinking in the nucleosome core. *J Mol Biol*, 368(4):1067–74, 2007.

[67] C. D. Schmid and P. Bucher. ChIP-Seq data reveal nucleosome architecture of human promoters. *Cell*, 131(5):831–2; author reply 832–3, 2007.

[68] R. D. Kornberg. Chromatin structure: a repeating unit of histones and DNA. *Science*, 184(139):868–71, 1974.

[69] R. D. Kornberg. Structure of chromatin. *Annu Rev Biochem*, 46:931–54, 1977.

[70] V. Ramakrishnan. Histone structure and the organization of the nucleosome. *Annu Rev Biophys Biomol Struct*, 26:83–112, 1997.

[71] R. D. Kornberg and Y. Lorch. Chromatin rules. *Nat Struct Mol Biol*, 14(11):986–8, 2007.

[72] A. Csordas. A proposal for a possible role of nucleosome positioning in the evolutionary adjustment of introns. *Int J Biochem*, 21(5):455–61, 1989.

[73] C. L. White, R. K. Suto, and K. Luger. Structure of the yeast nucleosome core particle reveals fundamental changes in internucleosome interactions. *EMBO J*, 20(18):5207–18, 2001.

[74] G. C. Yuan, Y. J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, 309(5734):626–30, 2005.

[75] P. M. Pryciak and H. E. Varmus. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell*, 69(5):769–80, 1992.

[76] C. Davey, R. Fraser, M. Smolle, M. W. Simmen, and J. Allan. Nucleosome positioning signals in the DNA sequence of the human and mouse H19 imprinting control regions. *J Mol Biol*, 325(5):873–87, 2003.

[77] E. A. Sekinger, Z. Moqtaderi, and K. Struhl. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol Cell*, 18(6):735–48, 2005.

[78] J. Zhou, J. Y. Fan, D. Rangasamy, and D. J. Tremethick. The nucleosome surface regulates chromatin compaction and couples it with transcriptional repression. *Nat Struct Mol Biol*, 14(11):1070–6, 2007.

[79] J. L. Gutierrez, M. Chandy, M. J. Carrozza, and J. L. Workman. Activation domains drive nucleosome eviction by SWI/SNF. *EMBO J*, 26(3):730–40, 2007.

[80] A. A. Travers and A. Klug. The bending of DNA in nucleosomes and its wider implications. *Philos Trans R Soc Lond B Biol Sci*, 317(1187):537–61, 1987.

[81] R. E. Dickerson and T. K. Chiu. Helix bending as a factor in protein/DNA recognition. *Biopolymers*, 44(4):361–403, 1997.

[82] T. J. Richmond and C. A. Davey. The structure of DNA in the nucleosome core. *Nature*, 423(6936):145–50, 2003.

[83] R. E. Dickerson. DNA bending: the prevalence of kinkiness and the virtues of normality. *Nucleic Acids Res*, 26(8):1906–26, 1998.

[84] K. Yanagi, G. G. Prive, and R. E. Dickerson. Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *J Mol Biol*, 217(1):201–14, 1991.

[85] M. A. El Hassan and C. R. Calladine. Two distinct modes of protein-induced bending in DNA. *J Mol Biol*, 282(2):331–43, 1998.

[86] V. B. Zhurkin, N. B. Ulyanov, A. A. Gorin, and R. L. Jernigan. Static and statistical bending of DNA evaluated by Monte Carlo simulations. *Proc Natl Acad Sci U S A*, 88(16):7046–50, 1991.

[87] W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock, and V. B. Zhurkin. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci U S A*, 95(19):11163–8, 1998.

[88] A. Stein and M. Bina. A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment. *Nucleic Acids Res*, 27(3):848–53, 1999.

[89] F. Thoma, L. W. Bergman, and R. T. Simpson. Nuclease digestion of circular TRP1ARS1 chromatin reveals positioned nucleosomes separated by nuclease-sensitive regions. *J Mol Biol*, 177(4):715–33, 1984.

[90] D. S. Pederson, F. Thoma, and R. T. Simpson. Core particle, fiber, and transcriptionally active chromatin structure. *Annu Rev Cell Biol*, 2:117–47, 1986.

[91] R. E. Dickerson, D. S. Goodsell, and S. Neidle. "...the tyranny of the lattice...". *Proc Natl Acad Sci U S A*, 91(9):3579–83, 1994.

[92] D. S. Goodsell, M. L. Kopka, D. Cascio, and R. E. Dickerson. Crystal structure of CATGGCCATG and its implications for A-tract bending models. *Proc Natl Acad Sci U S A*, 90(7):2930–4, 1993.

[93] I. Ioshikhes, A. Bolshoy, and E. N. Trifonov. Preferred positions of AA and TT dinucleotides in aligned nucleosomal DNA sequences. *J Biomol Struct Dyn*, 9(6):1111–7, 1992.

[94] M. J. Pazin, P. Bhargava, E. P. Geiduschek, and J. T. Kadonaga. Nucleosome mobility and the maintenance of nucleosome positioning. *Science*, 276(5313):809–12, 1997.

[95] R. T. Simpson. Nucleosome positioning: occurrence, mechanisms, and functional consequences. *Prog Nucleic Acid Res Mol Biol*, 40:143–84, 1991.

[96] B. Pina, D. Barettino, and M. Beato. Nucleosome positioning and regulated gene expression. *Oxf Surv Eukaryot Genes*, 7:83–117, 1991.

[97] P. T. Lowary and J. Widom. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *J Mol Biol*, 276(1):19–42, 1998.

[98] A. Thastrom, P. T. Lowary, H. R. Widlund, H. Cao, M. Kubista, and J. Widom. Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J Mol Biol*, 288(2):213–29, 1999.

[99] A. G. Fernandez and J. N. Anderson. Nucleosome positioning determinants. *J Mol Biol*, 371(3):649–68, 2007.

[100] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I. K. Moore, J. P. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–8, 2006.

[101] Z. Zhang, K. Shibahara, and B. Stillman. PCNA connects DNA replication to epigenetic inheritance in yeast. *Nature*, 408(6809):221–5, 2000.

[102] A. Loyola and G. Almouzni. Histone chaperones, a supporting role in the limelight. *Biochim Biophys Acta*, 1677(1-3):3–11, 2004.

[103] Y. Fan, T. Nikitina, E. M. Morin-Kensicki, J. Zhao, T. R. Magnuson, C. L. Woodcock, and A. I. Skoultchi. H1 linker histones are essential for mouse development and affect nucleosome spacing in vivo. *Mol Cell Biol*, 23(13):4559–72, 2003.

[104] C. L. Woodcock, A. I. Skoultchi, and Y. Fan. Role of linker histone in chromatin structure and function: H1 stoichiometry and nucleosome repeat length. *Chromosome Res*, 14(1):17–25, 2006.

[105] M. W. Adkins and J. K. Tyler. The histone chaperone Asf1p mediates global chromatin disassembly in vivo. *J Biol Chem*, 279(50):52069–74, 2004.

[106] S. Henikoff. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nat Rev Genet*, 9(1):15–26, 2008.

[107] A. Rufiange, P. E. Jacques, W. Bhat, F. Robert, and A. Nourani. Genome-wide replication-independent histone H3 exchange occurs predominantly at promoters and implicates H3 K56 acetylation and Asf1. *Mol Cell*, 27(3):393–405, 2007.

[108] V. Levitsky, O. Podkolodnaya, N. Kolchanov, and N. Podkolodny. Nucleosome formation potential of eukaryotic DNA: calculation and promoters analysis. *Bioinformatics*, 17(11):998–1010, 2001.

[109] V. Levitsky, O. Podkolodnaya, N. Kolchanov, and N. Podkolodny. Nucleosome formation potential of exons, introns, and Alu repeats. *Bioinformatics*, 17(11):1062–1010, 2001.

[110] H. E. Peckham, R. E. Thurman, Y. Fu, J. A. Stamatoyannopoulos, W. S. Noble, K. Struhl, and Z. Weng. Nucleosome positioning signals in genomic DNA. *Genome Res*, 17(8):1170–7, 2007.

[111] K. P. Nightingale, L. P. O'Neill, and B. M. Turner. Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Curr Opin Genet Dev*, 16(2):125–36, 2006.

[112] C. Bock and T. Lengauer. Computational epigenetics. *Bioinformatics*, 24(1):1–10, 2008.

[113] N. J. Francis and R. E. Kingston. Mechanisms of transcriptional memory. *Nat Rev Mol Cell Biol*, 2(6):409–21, 2001.

[114] B. M. Turner. Cellular memory and the histone code. *Cell*, 111(3):285–91, 2002.

[115] B. M. Turner. Memorable transcription. *Nat Cell Biol*, 5(5):390–3, 2003.

[116] W. Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425–32, 2007.

[117] R. Festenstein, S. N. Pagakis, K. Hiragami, D. Lyon, A. Verreault, B. Sekkali, and D. Kioussis. Modulation of heterochromatin protein 1 dynamics in primary Mammalian cells. *Science*, 299(5607):719–21, 2003.

[118] M. Rassoulzadegan, V. Grandjean, P. Gounon, and F. Cuzin. Inheritance of an epigenetic change in the mouse: a new role for RNA. *Biochem Soc Trans*, 35(Pt 3):623–5, 2007.

[119] T. Kouzarides. Chromatin modifications and their function. *Cell*, 128(4):693–705, 2007.

[120] B. M. Turner. Reading signals on the nucleosome with a new nomenclature for modified histones. *Nat Struct Mol Biol*, 12(2):110–2, 2005.

[121] Y. Zhang and D. Reinberg. Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes Dev*, 15(18):2343–60, 2001.

[122] R. J. Sims, 3rd, K. Nishioka, and D. Reinberg. Histone lysine methylation: a signature for chromatin function. *Trends Genet*, 19(11):629–39, 2003.

[123] T. Y. Roh, S. Cuddapah, and K. Zhao. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes Dev*, 19(5):542–52, 2005.

[124] S. J. Nowak, C. Y. Pai, and V. G. Corces. Protein phosphatase 2A activity affects histone H3 phosphorylation and transcription in Drosophila melanogaster. *Mol Cell Biol*, 23(17):6129–38, 2003.

[125] S. J. Nowak and V. G. Corces. Phosphorylation of histone H3: a balancing act between chromosome condensation and transcriptional activation. *Trends Genet*, 20(4):214–20, 2004.

[126] A. Shilatifard. Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. *Annu Rev Biochem*, 75:243–69, 2006.

[127] P. O. Hassa, S. S. Haenni, M. Elser, and M. O. Hottiger. Nuclear ADP-ribosylation reactions in mammalian cells: where are we today and where are we going? *Microbiol Mol Biol Rev*, 70(3):789–829, 2006.

[128] D. Nathan, K. Ingvarsdottir, D. E. Sterner, G. R. Bylebyl, M. Dokmanovic, J. A. Dorsey, K. A. Whelan, M. Krsmanovic, W. S. Lane, P. B. Meluh, E. S. Johnson, and S. L. Berger. Histone sumoylation is a negative regulator in Saccharomyces cerevisiae and shows dynamic interplay with positive-acting histone modifications. *Genes Dev*, 20(8):966–76, 2006.

[129] C. J. Nelson, H. Santos-Rosa, and T. Kouzarides. Proline isomerization of histone H3 regulates lysine methylation and gene expression. *Cell*, 126(5):905–16, 2006.

[130] J. A. Daniel, M. G. Pray-Grant, and P. A. Grant. Effector proteins for methylated histones: an expanding family. *Cell Cycle*, 4(7):919–26, 2005.

[131] M. D. Litt, M. Simpson, M. Gaszner, C. D. Allis, and G. Felsenfeld. Correlation between histone lysine methylation and developmental changes at the chicken beta-globin locus. *Science*, 293(5539):2453–5, 2001.

[132] V. Mutskov and G. Felsenfeld. Silencing of transgene transcription precedes methylation of promoter DNA and histone H3 lysine 9. *EMBO J*, 23(1):138–49, 2004.

[133] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, 2007.

[134] D. K. Pokholok, C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, E. Herbolsheimer, J. Zeitlinger, F. Lewitter, D. K. Gifford, and R. A. Young. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, 122(4):517–27, 2005.

[135] M. F. Dion, S. J. Altschuler, L. F. Wu, and O. J. Rando. Genomic characterization reveals a simple histone H4 acetylation code. *Proc Natl Acad Sci U S A*, 102(15):5501–6, 2005.

[136] T. S. Mikkelsen, M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S.

Lander, and B. E. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–60, 2007.

[137] T. Agalioti, S. Lomvardas, B. Parekh, J. Yie, T. Maniatis, and D. Thanos. Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. *Cell*, 103(4):667–78, 2000.

[138] T. Agalioti, G. Chen, and D. Thanos. Deciphering the transcriptional histone acetylation code for a human gene. *Cell*, 111(3):381–92, 2002.

[139] R. Metivier, G. Penot, M. R. Hubner, G. Reid, H. Brand, M. Kos, and F. Gannon. Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell*, 115(6):751–63, 2003.

[140] M. J. Carrozza, R. T. Utley, J. L. Workman, and J. Cote. The diverse functions of histone acetyltransferase complexes. *Trends Genet*, 19(6):321–9, 2003.

[141] D. Schubeler, D. M. MacAlpine, D. Scalzo, C. Wirbelauer, C. Kooperberg, F. van Leeuwen, D. E. Gottschling, L. P. O'Neill, B. M. Turner, J. Delrow, S. P. Bell, and M. Groudine. The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote. *Genes Dev*, 18(11):1263–71, 2004.

[142] B. E. Bernstein, E. L. Humphrey, C. L. Liu, and S. L. Schreiber. The use of chromatin immunoprecipitation assays in genome-wide analyses of histone modifications. *Methods Enzymol*, 376:349–60, 2004.

[143] D. J. Huebert, M. Kamal, A. O'Donovan, and B. E. Bernstein. Genome-wide analysis of histone modifications by ChIP-on-chip. *Methods*, 40(4):365–9, 2006.

[144] B. E. Bernstein, M. Kamal, K. Lindblad-Toh, S. Bekiranov, D. K. Bailey, D. J. Huebert, S. McMahon, E. K. Karlsson, E. J. Kulbokas, 3rd, T. R. Gingeras, S. L. Schreiber, and E. S. Lander. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2):169–81, 2005.

[145] B. Macek, L. F. Waanders, J. V. Olsen, and M. Mann. Top-down protein sequencing and MS3 on a hybrid linear quadrupole ion trap-orbitrap mass spectrometer. *Mol Cell Proteomics*, 5(5):949–58, 2006.

[146] S. Henikoff and K. Ahmad. Assembly of variant histones into chromatin. *Annu Rev Cell Dev Biol*, 21:133–53, 2005.

[147] K. Ahmad and S. Henikoff. Histone H3 variants specify modes of chromatin assembly. *Proc Natl Acad Sci U S A*, 99 Suppl 4:16477–84, 2002.

[148] K. Ahmad and S. Henikoff. The histone variant H3.3 marks active chromatin by replication-independent nucleosome assembly. *Mol Cell*, 9(6):1191–200, 2002.

[149] Y. Mito, J. G. Henikoff, and S. Henikoff. Genome-scale profiling of histone H3.3 replacement patterns. *Nat Genet*, 37(10):1090–7, 2005.

[150] Y. Mito, J. G. Henikoff, and S. Henikoff. Histone replacement marks the boundaries of cis-regulatory domains. *Science*, 315(5817):1408–11, 2007.

[151] C. Jin and G. Felsenfeld. Distribution of histone H3.3 in hematopoietic cell lineages. *Proc Natl Acad Sci U S A*, 103(3):574–9, 2006.

[152] A. J. Antczak, T. Tsubota, P. D. Kaufman, and J. M. Berger. Structure of the yeast histone H3-ASF1 interaction: implications for chaperone mechanism, species-specific interactions, and epigenetics. *BMC Struct Biol*, 6:26, 2006.

[153] P. J. Sabo, R. Humbert, M. Hawrylycz, J. C. Wallace, M. O. Dorschner, M. McArthur, and J. A. Stamatoyannopoulos. Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries. *Proc Natl Acad Sci U S A*, 101(13):4537–42, 2004.

[154] H. Reinke and W. Horz. Histones are first hyperacetylated and then lose contact with the activated PHO5 promoter. *Mol Cell*, 11(6):1599–607, 2003.

[155] C. M. Chow, A. Georgiou, H. Szutorisz, A. Maia e Silva, A. Pombo, I. Barahona, E. Dargelos, C. Canzonetta, and N. Dillon. Variant histone H3.3 marks promoters of transcriptionally active genes during mammalian cell division. *EMBO Rep*, 6(4):354–60, 2005.

[156] B. E. Schwartz and K. Ahmad. Transcriptional activation triggers deposition and removal of the histone variant H3.3. *Genes Dev*, 19(7):804–14, 2005.

[157] U. J. Schermer, P. Korber, and W. Horz. Histones are incorporated in trans during reassembly of the yeast PHO5 promoter. *Mol Cell*, 19(2):279–85, 2005.

[158] M. F. Dion, T. Kaplan, M. Kim, S. Buratowski, N. Friedman, and O. J. Rando. Dynamics of replication-independent histone turnover in budding yeast. *Science*, 315(5817):1405–8, 2007.

[159] A. Jamai, R. M. Imoberdorf, and M. Strubin. Continuous histone H2B and transcription-dependent histone H3 exchange in yeast cells outside of replication. *Mol Cell*, 25(3):345–55, 2007.

[160] C. Moorman, L. V. Sun, J. Wang, E. de Wit, W. Talhout, L. D. Ward, F. Greil, X. J. Lu, K. P. White, H. J. Bussemaker, and B. van Steensel. Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. *Proc Natl Acad Sci U S A*, 103(32):12027–32, 2006.

[161] H. S. Malik and S. Henikoff. Phylogenomics of the nucleosome. *Nat Struct Biol*, 10(11):882–91, 2003.

[162] B. P. Chadwick and H. F. Willard. A novel chromatin protein, distantly related to histone H2A, is largely excluded from the inactive X chromosome. *J Cell Biol*, 152(2):375–84, 2001.

[163] S. Henikoff, T. Furuyama, and K. Ahmad. Histone variants, nucleosome assembly and epigenetic inheritance. *Trends Genet*, 20(7):320–6, 2004.

[164] B. E. Bernstein, A. Meissner, and E. S. Lander. The mammalian epigenome. *Cell*, 128(4):669–81, 2007.

[165] A. Bird. DNA methylation patterns and epigenetic memory. *Genes Dev*, 16(1):6–21, 2002.

[166] T. Ushijima, N. Watanabe, E. Okochi, A. Kaneda, T. Sugimura, and K. Miyamoto. Fidelity of the methylation pattern and its variation in the genome. *Genome Res*, 13(5):868–74, 2003.

[167] J. W. Drake, B. Charlesworth, D. Charlesworth, and J. F. Crow. Rates of spontaneous mutation. *Genetics*, 148(4):1667–86, 1998.

[168] S. W. Chan, I. R. Henderson, and S. E. Jacobsen. Gardening the genome: DNA methylation in Arabidopsis thaliana. *Nat Rev Genet*, 6(5):351–60, 2005.

[169] T. R. Haines, D. I. Rodenhiser, and P. J. Ainsworth. Allele-specific non-CpG methylation of the Nf1 gene during early mouse development. *Dev Biol*, 240(2):585–98, 2001.

[170] B. H. Ramsahoye, D. Biniszkiewicz, F. Lyko, V. Clark, A. P. Bird, and R. Jaenisch. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proc Natl Acad Sci U S A*, 97(10):5237–42, 2000.

[171] A. Bird. Perceptions of epigenetics. *Nature*, 447(7143):396–8, 2007.

[172] T. Jenuwein and C. D. Allis. Translating the histone code. *Science*, 293(5532):1074–80, 2001.

[173] H. D. Morgan, F. Santos, K. Green, W. Dean, and W. Reik. Epigenetic reprogramming in mammals. *Hum Mol Genet*, 14 Spec No 1:R47–58, 2005.

[174] T. Kaneko-Ishino, T. Kohda, and F. Ishino. The regulation and biological significance of genomic imprinting in mammals. *J Biochem*, 133(6):699–711, 2003.

[175] R. Feil and F. Berger. Convergent evolution of genomic imprinting in plants and mammals. *Trends Genet*, 23(4):192–9, 2007.

[176] G. Kelsey and W. Reik. Imprint switch mechanism indicated by mutations in Prader-Willi and Angelman syndromes. *Bioessays*, 19(5):361–5, 1997.

[177] W. Reik and E. R. Maher. Imprinting in clusters: lessons from Beckwith-Wiedemann syndrome. *Trends Genet*, 13(8):330–4, 1997.

[178] C. B. Santos-Reboucas and M. M. Pimentel. Implication of abnormal epigenetic patterns for human diseases. *Eur J Hum Genet*, 15(1):10–7, 2007.

[179] M. Guo, M. G. House, H. Suzuki, Y. Ye, M. V. Brock, F. Lu, Z. Liu, A. K. Rustgi, and J. G. Herman. Epigenetic silencing of CDX2 is a feature of squamous esophageal cancer. *Int J Cancer*, 121(6):1219–26, 2007.

[180] F. Fuks. DNA methylation and histone modifications: teaming up to silence genes. *Curr Opin Genet Dev*, 15(5):490–5, 2005.

[181] A. M. Lindroth, D. Shultis, Z. Jasencakova, J. Fuchs, L. Johnson, D. Schubert, D. Patnaik, S. Pradhan, J. Goodrich, I. Schubert, T. Jenuwein, S. Khorasanizadeh, and S. E. Jacobsen. Dual histone H3 methylation marks at lysines 9 and 27 required for interaction with CHROMOMETHYLASE3. *EMBO J*, 23(21):4286–96, 2004.

[182] O. Mathieu, A. V. Probst, and J. Paszkowski. Distinct regulation of histone H3 methylation at lysines 27 and 9 by CpG methylation in Arabidopsis. *EMBO J*, 24(15):2783–91, 2005.

[183] Y. Habu, O. Mathieu, M. Tariq, A. V. Probst, C. Smathajitt, T. Zhu, and J. Paszkowski. Epigenetic regulation of transcription in intermediate heterochromatin. *EMBO Rep*, 7(12):1279–84, 2006.

[184] E. Vire, C. Brenner, R. Deplus, L. Blanchon, M. Fraga, C. Didelot, L. Morey, A. Van Eynde, D. Bernard, J. M. Vanderwinden, M. Bollen, M. Esteller, L. Di Croce, Y. de Launoit, and F. Fuks. The Polycomb group protein EZH2 directly controls DNA methylation. *Nature*, 439(7078):871–4, 2006.

[185] M. Nekrasov, T. Klymenko, S. Fraterman, B. Papp, K. Oktaba, T. Kocher, A. Cohen, H. G. Stunnenberg, M. Wilm, and J. Muller. Pcl-PRC2 is needed to generate high levels of H3-K27 trimethylation at Polycomb target genes. *EMBO J*, 26(18):4078–88, 2007.

[186] Y. B. Schwartz, T. G. Kahn, D. A. Nix, X. Y. Li, R. Bourgon, M. Biggin, and V. Pirrotta. Genome-wide analysis of Polycomb targets in Drosophila melanogaster. *Nat Genet*, 38(6):700–5, 2006.

[187] Y. B. Schwartz and V. Pirrotta. Polycomb silencing mechanisms and the management of genomic programmes. *Nat Rev Genet*, 8(1):9–22, 2007.

[188] H. W. Brock and C. L. Fisher. Maintenance of gene expression patterns. *Dev Dyn*, 232(3):633–55, 2005.

[189] J. Secombe, L. Li, L. Carlos, and R. N. Eisenman. The Trithorax group protein Lid is a trimethyl histone H3K4 demethylase required for dMyc-induced cell growth. *Genes Dev*, 21(5):537–51, 2007.

[190] J. Secombe and R. N. Eisenman. The function and regulation of the JARID1 family of histone H3 lysine 4 demethylases: the Myc connection. *Cell Cycle*, 6(11):1324–8, 2007.

[191] A. Sparmann and M. van Lohuizen. Polycomb silencers control cell fate, development and cancer. *Nat Rev Cancer*, 6(11):846–56, 2006.

[192] Y. M. Moshkin, J. A. Armstrong, R. K. Maeda, J. W. Tamkun, P. Verrijzer, J. A. Kennison, and F. Karch. Histone chaperone ASF1 cooperates with the Brahma chromatin-remodelling machinery. *Genes Dev*, 16(20):2621–6, 2002.

[193] B. Papp and J. Muller. Histone trimethylation and the maintenance of transcriptional ON and OFF states by trxG and PcG proteins. *Genes Dev*, 20(15):2041–54, 2006.

[194] L. Hennig, R. Bouveret, and W. Gruissem. MSI1-like proteins: an escort service for chromatin assembly and remodeling complexes. *Trends Cell Biol*, 15(6):295–302, 2005.

[195] C. Martin and Y. Zhang. Mechanisms of epigenetic inheritance. *Curr Opin Cell Biol*, 19(3):266–72, 2007.

[196] R. J. Klose and A. P. Bird. Genomic DNA methylation: the mark and its mediators. *Trends Biochem Sci*, 31(2):89–97, 2006.

[197] E. Valls, S. Sanchez-Molina, and M. A. Martinez-Balbas. Role of histone modifications in marking and activating genes through mitosis. *J Biol Chem*, 280(52):42592–600, 2005.

[198] S. Henikoff. Histone modifications: combinatorial complexity or cumulative simplicity? *Proc Natl Acad Sci U S A*, 102(15):5308–9, 2005.

[199] Y. Nakatani, D. Ray-Gallet, J. P. Quivy, H. Tagami, and G. Almouzni. Two distinct nucleosome assembly pathways: dependent or independent of DNA synthesis promoted by histone H3.1 and H3.3 complexes. *Cold Spring Harb Symp Quant Biol*, 69:273–80, 2004.

[200] H. Tagami, D. Ray-Gallet, G. Almouzni, and Y. Nakatani. Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis. *Cell*, 116(1):51–61, 2004.

[201] H. Weintraub, A. Worcel, and B. Alberts. A model for chromatin based upon two symmetrically paired half-nucleosomes. *Cell*, 9(3):409–17, 1976.

[202] A. T. Annunziato. Split decision: what happens to nucleosomes during DNA replication? *J Biol Chem*, 280(13):12065–8, 2005.

[203] M. Ptashne. On the use of the word 'epigenetic'. *Curr Biol*, 17(7):R233–6, 2007.

[204] R. Natsume, M. Eitoku, Y. Akai, N. Sano, M. Horikoshi, and T. Senda. Structure and function of the histone chaperone CIA/ASF1 complexed with histones H3 and H4. *Nature*, 446(7133):338–41, 2007.

[205] A. K. Nagaich, D. A. Walker, R. Wolford, and G. L. Hager. Rapid periodic binding and displacement of the glucocorticoid receptor during chromatin remodeling. *Mol Cell*, 14(2):163–74, 2004.

[206] J. G. McNally, W. G. Muller, D. Walker, R. Wolford, and G. L. Hager. The glucocorticoid receptor: rapid exchange with regulatory sites in living cells. *Science*, 287(5456):1262–5, 2000.

[207] D. Bosisio, I. Marazzi, A. Agresti, N. Shimizu, M. E. Bianchi, and G. Natoli. A hyperdynamic equilibrium between promoter-bound and nucleoplasmic dimers controls NF-kappaB-dependent gene activity. *EMBO J*, 25(4):798–810, 2006.

[208] M. J. Buck and J. D. Lieb. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–60, 2004.

[209] I. M. Wilson, J. J. Davies, M. Weber, C. J. Brown, C. E. Alvarez, C. MacAulay, D. Schubeler, and W. L. Lam. Epigenomics: mapping the methylome. *Cell Cycle*, 5(2):155–8, 2006.

[210] P. Hajkova, S. Erhardt, N. Lane, T. Haaf, O. El-Maarri, W. Reik, J. Walter, and M. A. Surani. Epigenetic reprogramming in mouse primordial germ cells. *Mech Dev*, 117(1-2):15–23, 2002.

[211] P. Hajkova, O. el Maarri, S. Engemann, J. Oswald, A. Olek, and J. Walter. DNA-methylation analysis by the bisulfite-assisted genomic sequencing method. *Methods Mol Biol*, 200:143–54, 2002.

[212] T. E. Royce, J. S. Rozowsky, P. Bertone, M. Samanta, V. Stolc, S. Weissman, M. Snyder, and M. Gerstein. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet*, 21(8):466–75, 2005.

[213] S. Cawley, S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, and T. R. Gingeras. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, 116(4):499–509, 2004.

[214] W. Li, C. A. Meyer, and X. S. Liu. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, 21 Suppl 1:i274–82, 2005.

[215] W. E. Johnson, W. Li, C. A. Meyer, R. Gottardo, J. S. Carroll, M. Brown, and X. S. Liu. Model-based analysis of tiling-arrays for ChIP-chip. *Proc Natl Acad Sci U S A*, 103(33):12457–62, 2006.

[216] J. S. Song, W. E. Johnson, X. Zhu, X. Zhang, W. Li, A. K. Manrai, J. S. Liu, R. Chen, and X. S. Liu. Model-based analysis of two-color arrays (MA2C). *Genome Biol*, 8(8):R178, 2007.

[217] H. Ji and W. H. Wong. TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, 21(18):3629–36, 2005.

[218] M. J. Buck, A. B. Nobel, and J. D. Lieb. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol*, 6(11):R97, 2005.

[219] J. Toedling, O. Sklyar, and W. Huber. Ringo–an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics*, 8:221, 2007.

[220] J. Du, J. S. Rozowsky, J. O. Korbel, Z. D. Zhang, T. E. Royce, M. H. Schultz, M. Snyder, and M. Gerstein. A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics*, 22(24):3016–24, 2006.

[221] W. J. Kent. BLAT–the BLAST-like alignment tool. *Genome Res*, 12(4):656–64, 2002.

[222] V. B. Bajic, S. L. Tan, Y. Suzuki, and S. Sugano. Promoter prediction analysis on the whole human genome. *Nat Biotechnol*, 22(11):1467–73, 2004.

[223] A. D. Smith, P. Sumazin, and M. Q. Zhang. Tissue-specific regulatory elements in mammalian promoters. *Mol Syst Biol*, 3:73, 2007.

[224] Tim Van den Bulcke, Karen Lemmens, Yves Van de Peer, and Kathleen Marchal. Inferring Transcriptional Networks by Mining 'O'mics Data. *Current Bioinformatics*, 1(4):301–313, 2006.

[225] F. Antequera. Structure, function and evolution of CpG island promoters. *Cell Mol Life Sci*, 60(8):1647–58, 2003.

[226] D. F. Burke, C. L. Worth, E. M. Priego, T. Cheng, L. J. Smink, J. A. Todd, and T. L. Blundell. Genome bioinformatic analysis of nonsynonymous SNPs. *BMC Bioinformatics*, 8:301, 2007.

[227] C. Bock, M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer, and J. Walter. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet*, 2(3):e26, 2006.

[228] R. Das, N. Dimitrova, Z. Xuan, R. A. Rollins, F. Haghighi, J. R. Edwards, J. Ju, T. H. Bestor, and M. Q. Zhang. Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci U S A*, 103(28):10713–6, 2006.

[229] F. A. Feltus, E. K. Lee, J. F. Costello, C. Plass, and P. M. Vertino. Predicting aberrant CpG island methylation. *Proc Natl Acad Sci U S A*, 100(21):12253–8, 2003.

[230] F. A. Feltus, E. K. Lee, J. F. Costello, C. Plass, and P. M. Vertino. DNA motifs associated with aberrant CpG island methylation. *Genomics*, 87(5):572–9, 2006.

[231] A. Li, Y. Xue, C. Jin, M. Wang, and X. Yao. Prediction of N^{ε}-acetylation on internal lysines implemented in Bayesian Discriminant Method. *Biochem Biophys Res Commun*, 350(4):818–24, 2006.

[232] P. Taghavi and M. van Lohuizen. Developmental biology: two paths to silence merge. *Nature*, 439(7078):794–5, 2006.

Chapter 2

Specific Motif Context in HIV integration target sequences

2.1 Introduction

In 2002, Schroeder *et al.* demonstrated conclusively for the first time that HIV integration is a non-random event (1). Multiple early reports had indicated that HIV integration is indeed non-random, however, none of them provided compelling evidence with adequate controls and more importantly, using data from integrations spanning the entire genome. To discuss the genesis of project, we will first discuss this particular article. Another important study comparing the integration sites of various retro-viruses by Mitchell *et al.* will be also discussed (2). Previous studies had demonstrated an overlap between the target sites selected by the retrotransposons in yeast and the retroviruses in higher animals suggesting similarities in their targeting mechanisms (3). The study of retroviral integration sites is thus important not only to understand the biology of retroviruses, but also to understand the contribution of host factors and in turn, for designing better strategies for gene therapy.

2.2 Patho-physiology of HIV infection

The Human Immuno Deficiency Virus is a RNA[1] virus. It belongs to the superfamily of Retrotranscribing viruses, family Retroviridae (RNA as genomic material which is reverse transcribed into DNA), sub-family Orthoretrovirinae, clade Lentivirus, and specifically the primate lentivirus group. Most of these viruses have fastidious requirement of specific host/

[1]Possesses strands of sense RNA as genetic material

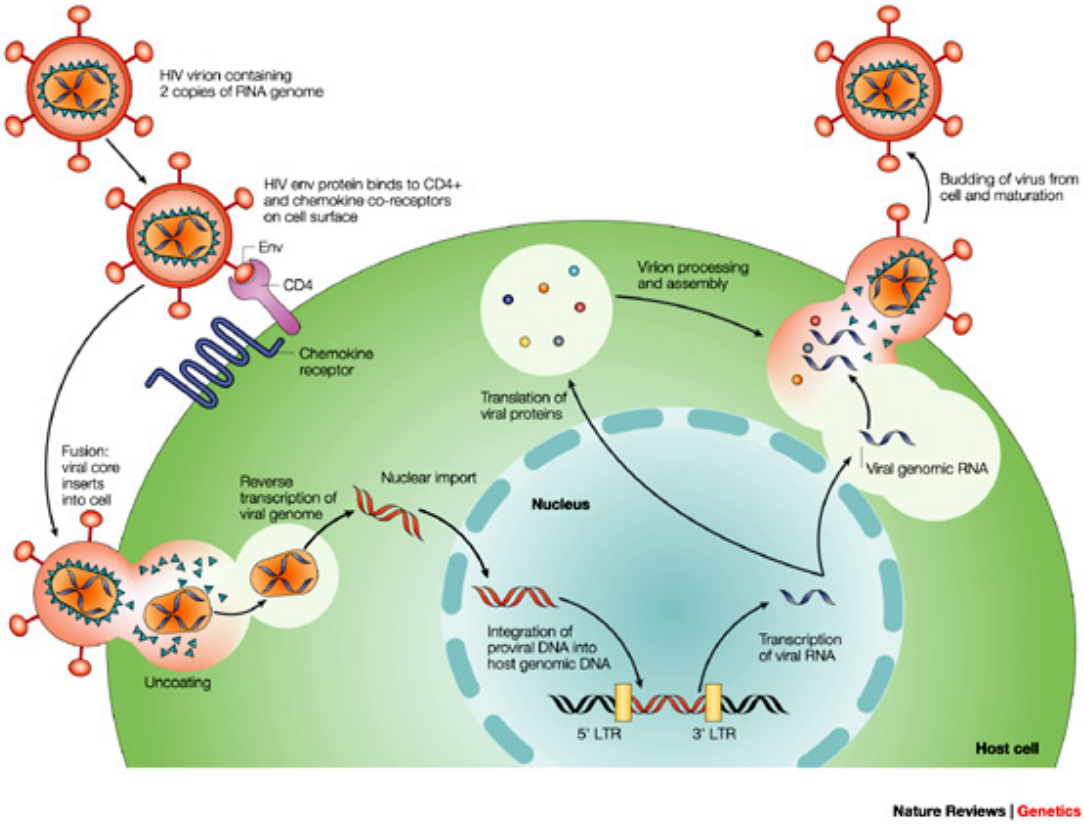


Figure 2.1: **Schematic representation of HIV lifecycle:** A schematic cartoon of HIV lifecycle (adapted from (4)).

cell-type. Many host and virus proteins play role in this species restriction (5). This virus is the primary causative agent of the acquired immuno deficiency syndrome (AIDS). The HIV in particular attacks the CD4+ve cells and destroys the ability of the affected individual to mount a immune response (6).

Following fusion of the virus with the host cell, the genetic material of the virus is released in the cytoplasm and undergoes reverse transcription into DNA and is simultaneously used as mRNA to produce virus specific proteins necessary for completion of its lifecycle (7). One of the virus specific proteins produced is the enzyme reverse transcriptase which is necessary to catalyze this conversion of viral RNA into cDNA. Specific viral proteins and host proteins associate with the viral cDNA forming the pre-integration complex (PIC) (8). The PIC migrates to the nucleus and brings about integration of the viral cDNA with the host genome (9). Such an 'integrated' cDNA may persist in the host for number of years asymptomatically.

Activation of the host cells results in the transcription of viral DNA into messenger RNA (mRNA), which is then translated into viral proteins (10). The exact mechanisms of the conversion of latent virus to cytolytic phase are not clear (11, see for a review). The new viral RNA forms the genetic material for the next generation of viruses. The viral RNA and viral proteins assemble at the cell membrane and mature into a new virion. Amongst the viral proteins the HIV protease is required to process other HIV proteins into their functional forms. Following assembly at the cell surface, the virus buds from the cell and is released to infect another cell. Unless the HIV life-cycle is interrupted by treatment with anti-retroviral agents, the virus infection spreads throughout the body and results in the destruction of the body's immune system, which leads to the AIDS.

2.3 Brief history of HIV infection

Beginning of an end HIV begins its infection of a susceptible host cell by binding to the CD4 receptor on the host cell. Recently it has been discovered that CD4+ cells are susceptible to recurrent infection by HIV (6). CD4 is present on the surface of many lymphocytes, which are a critical part of the bodys immune system. Recent evidence indicates that a co-receptor is needed for HIV to enter the cell (12).

Fusion Fusion of the virus particle with the host cell is the first step in HIV infection. As mentioned earlier CD4+ is the major receptor for the HIV. However, many cell surface proteins also act as receptors for various infective viruses. In an excellent review by Bour *et al.* an exhaustive list of cellular receptors of the retroviruses is provided (7). Some viruses, e.g. ASLV-A, MuLV-E, BLV, etc. use proteins on the host cell surface, that normally function in the cell as amino-acid transporters. The gp120 is a protein encoded by the virus and is part of its capsid. This protein acts as the primary viral partner in the starting of the fusion reaction. Fusion is thus a very important step in the virus lifecycle.

Role of the cell cycle Cell cycle effects on the virus lifecycle are virus specific. Gammaretroviruses[2] require mitosis for pro-viral integration whereas lentiviruses are able to repli-

[2]The genome is dimeric; not segmented and consists of a single molecule of linear, positive-sense, single-stranded RNA.

cate in post-mitotic non-dividing cells (13). Resting cells such as naïve resting T lymphocytes from peripheral blood cannot be productively infected by retroviruses, including lentiviruses, but the molecular basis of this restriction remains poorly understood (5). Initial studies in retroviral infections had found that the synthesis of virus specific RNA was independent of the cell cycle (14). It has been shown that there is efficient accumulation of nuclear forms of Avian Sarcoma Virus DNA in γ-irradiation arrested cells (15). Katz *et al.* have showed that the cell cycle plays an important role in the ability of the HIV virus to infect the cell (13). Majority of cells of an animal host are not progressing rapidly through the cell cycle, and such a cellular environment appears to be suboptimal for replication of all retroviruses. Moreover, it has also been demonstrated that HIV-1 integrates into the genomes of in vitro inoculated resting CD4+ T cells that have not received activating stimuli and have not entered cell cycle stage G(1b) (16). More recently it has been shown that the early stages of the HIV life cycle are inefficient in post-stimulated CD4+ cells and that efficient replication cannot be induced by subsequent activation (17).

Role of the host proteins The integration of the HIV pre-integration complex (PIC) is a very complex process. Many cellular and virus encoded proteins play important role in this process. It has been known that HMG-I(Y) is part of the pre-integration complex (18). Most notably the SWI-SNF complex (19), and the LEDGF/p75 (20) are also part of this complex. HIV as a retrovirus is a very sophisticated and efficient instrument of payload delivery. But it does not have all the required machinery for the sustenance and reproduction and therefore depends on many host proteins for the same. Apart from the proteins mentioned above, many other cellular proteins play a role in infectivity and pathogenicity of the HIV in the host cell. Some other well documented proteins are barrier-to-autointegration factor (BAF), Ku, lamina-associated polypeptide 2a (LAP2a), (21, and references therein).

2.4 What is known about HIV integration target site selection?

Integration of retroviral cDNA into the host cell chromosome is an essential obligatory step in its replication virus lifecycle (22). Integration of the virus or its PIC on naked DNA

is considered non-specific (1). However, inside a cell, the PIC is presented with chromatin and not the naked DNA. Specific sequences in the viral genome that are necessary for the integration into the host genome have been characterized (23). This process is catalyzed by the retroviral integrase protein, which is conserved among retroviruses and retrotransposons (24, 4). Integrase is important part of the functional PIC (24). While the PIC is capable of directing integration of the viral cDNA at any chromosomal location, different retroviruses have clear preferences for integration in or near particular chromosomal features (25, 26, 27, 28). Lewinski *et al.* performed a comparative study in this regard and demonstrated that the HIV Gag plays an important role in targeting of the pro-virus to the genomic DNA (29). HIV has been shown to preferentially integrate in the vicinity of the transcription start sequences (2, 30). There is also evidence to show that in untreated HIV infected patients, the HIV integration occurs preferentially within genes (31). Similarly, there are specific regions of the genome which are avoided by the virus for integration, e.g. the centromeric alphoid repeats (32). In vitro studies have shown that HIV integration targets DNA with protein induced bending, suggesting that the site selection requires more information than the sequence itself (33). The mechanism of such a specific site selection has been an enigma. Only recently some observations have made better understanding of this aspect of HIV integration possible (34, 35). Holman & Coffin, have demonstrated that there is a symmetrical base preference around the provirus integration site (34). Jhonson & Levy have shown that matrix attachment regions (MARs) influence the target site selection by the HIV PIC (35). The primary DNA sequence features have been implicated in the target site selection by the HIV-PIC in vivo(36). The observations of Leclercq *et al.* and Jhonson & Levy are significant for further discussion (Section 2.5, on page 53), they say that if most or all the regions of the genome appear to be accessible to HTLV-1 integration, local DNA curvature seems to confer a kinetic advantage for both in vitro and in vivo HTLV-1 integration (35, 36).

2.5 HIV integration hotspots are also rich in SATB1 binding sequences

In the year 2002 Schröder *et al.* published an article that showed that HIV integration is not random but occurs at specific spots in the genome (1). At the same time in our lab attempts were on to find genome-wide targets for the chromatin modulator SATB1. To locate the gnomic locations where SATB1 binding sequences are present, a well known technique of Chromatin Immuno-precipitation (ChIP) was used (37). Briefly, the chromatin in the cell (monolayer or suspension culture) was crosslinked using formaldehyde. The crosslinked chromatin was extracted, and immunoprecipitated using α-SATB1 antibody. The cross links were reversed and the genomic DNA was extracted from the immunoprecipitated chromatin. To monitor presence of the genomic region of interest PCR is then carried out using specific primers. Alternatively, for identification of novel sites DNA from the immunoprecipitated chromatin is purified, cloned and sequenced. Most astonishingly, the sequences thus obtained were from a reported HIV integration hot spot on chromosome 11 as reported by Schröder *et al.* (1). SATB1 is global chromatin organizer and transcription factor. One of its functions is to anchor the genomic DNA to the nuclear matrix, by binding to the MARs (38). When seen in context of Section 2.4, we hypothesized presence of a link between the genomic DNA sequence and the target site selection by the HIV-PIC. As a T-lineage-enriched global chromatin organizer, we proposed that SATB1 could be the host factor that is targeted by the PIC and therefore could dictate the integration site choice.

2.6 HIV integration occurs at specific location in the genome

HIV prefers certain regions in the chromosomes for its integration over others. To unequivocally establish this, Schroeder *et al.* used a very elegant procedure to demonstrate the non-randomness of HIV integration (1). Briefly, SupT1 cells were infected with HIV in vitro. Genomic DNA was obtained from these infected cells. This genomic DNA was subjected restriction digestion. These digested fragments were used as template for PCR. One of the primers' corresponded to the LTR of the integrated virus, and the other primer was directed at the restriction enzyme site used to digest the genomic DNA. Thus, they amplified selec-

tively only those regions of the genome that had the virus integrated. A similar exercise was carried out by exposing the purified (naked) genomic DNA to the PIC in vitro. Both the sets of sequences were submitted to NCBI and were available in the public domain for download. This dataset provided the starting point for our analysis.

2.7 Alu-like motifs are enriched in sequences flanking the reported HIV-1 integration sequences.

2.7.1 Preliminary Sequence Analysis

We initially performed a gapped alignment of sets of cloned integration sequences deposited in NCBI using `ClustalX` (39). Multiple sequence alignments of in vivo integration sequences revealed a pattern. We found that these sequences share extended homologous regions which were spread across the lengths of the sequences. The sequence similarity appeared to be present in 'chunks of similar sequences in nearly all the sequences taken for alignment.

Initially, the HIV-1 integration sites were retrieved through the in vivo experiments referred in Section 2.4, however, it resulted a sequence set of varying length. The length of the reported sequences flanking integration sites sequences in the `GenBank` varies from 26 base pairs to over 1700 base pairs. More importantly, the manner in which the integration sequences were cloned represented only the $3'$ half of the integration site (1). We therefore `BLAST`ed (40) each of the reported integration sites onto the reference sequence of the human genome, and acquired the $5'$ and $3'$ flanking regions such that for each integration site a length of 2 kb sequence was obtained, for maintaining uniformity in analysis. A set of 429 sequences each with size 2 kb were used for further analysis. Multiple sequence alignments (MSA) were performed for such a data set which revealed regions of homologies among multiple sequences as in case of the original in vivo experiments as seen in Figure 2.3. The scheme describing the preparation of the data is depicted in Figure 2.2.

A single unrooted phylogenetic tree was plotted for these sequences (genomic sequences flanking the mapped HIV integration sites) as shown in Figure 2.4. The phylogenetic tree showed one major branch of related sequences, which comprised of more than 60% of the sequences (Figure 2.4). As a control data set, 450 sequences were simulated using the first-

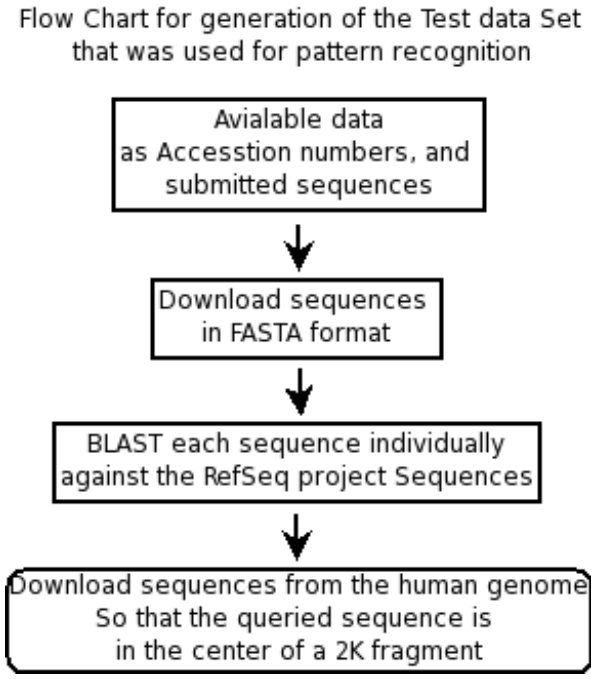


Figure 2.2: Scheme for Data Preparation.

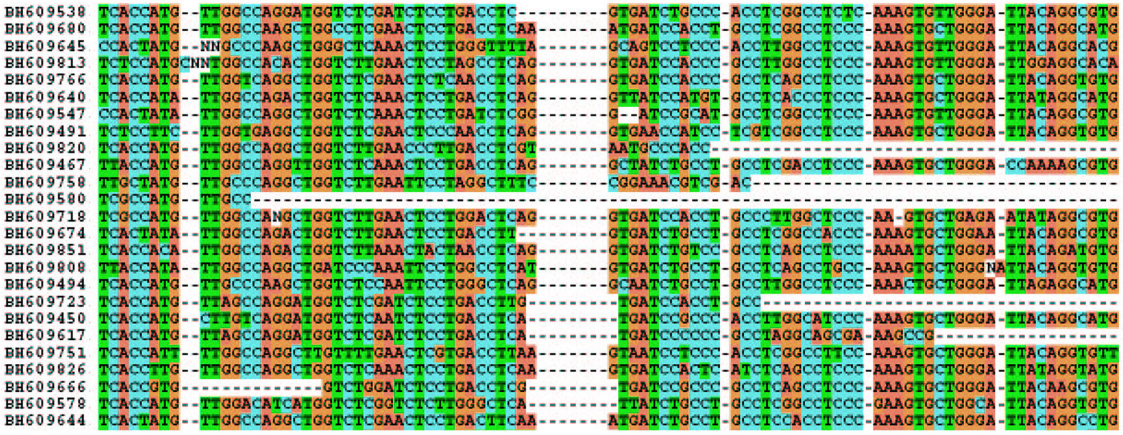


Figure 2.3: **Regions flanking the integration sequences show specific conserved regions:** MSA of the flanking regions of the integration sites was visualized as sequences. The neat 'chunks' of similarity can be clearly seen, showing that there are regions around the integration sequences that are nearly identical at all the integration sites.

order Markov chain simulator considering the random transition of bases amongst themselves. We also used fifth order Markov chain simulator to generate a set of randomly picked up sequences from the human genome. We refer to the former control sequence data set as *simulated random* and the latter as *random*. The unrooted phylogenetic trees for both these random sequence data sets displayed virtually no relatedness amongst individual sequences (Figures 2.5 and 2.6). Each of the Figures 2.5 and 2.6, actually shows unrooted tree of 450 sequences, such that each individual line, denotes a sequence. Figure 2.4 prompted us to look for presence of motifs if any in the given set of sequences (See Section 2.4).

Furthermore, the sequences were compared against the Reference Human Genome (Build 35 version 3) database. The distribution of the integration sites in the genome (only the real invivo sequences as mentioned by Schröeder *et al.* (1) were taken for this analysis) is shown in the Figures 2.7 and 2.8. It can be clearly seen from Figure 2.7 that the number of integration sites per chromosome corresponds to the Gene Density (number of genes per mega base pair of genomic DNA) of the chromosome. In Figure 2.7 the red impluses denote the number of integration sites in each chromosome (the abscissa). The black line denotes the Gene Density of each chromosome (plotted on the second y axis). Also, it is clearly seen there that the maximum number of integration events occur in chromosome number 19, which is the most gene-rich chromosome in the human genome.

The same relation also exists with respect to absolute number of genes per chromosome (Figure 2.9). Further there is no direct relation ship between the length of the chromosome and number integration events scored per chromosome. No integration events were seen in the Y chromosome. This correlation is further illustrated in Figure 2.10. It can be seen that there is a clear positive correlation between the Gene Density and the retroviral integration events. The correlation co-efficient R-squared is 0.79 (p-value $\ll 0.001$).

Although there is a positive correlation between the length of a chromosome and number of integration events, it is a weak correlation (R-squared = 0.43, and p-value = 0.03), as can be seen in Figure 2.11 and Figure 2.12. In Figure 2.12, the red bars denote length of chromosomes and are plotted on the Y-axis to the left, the black line represents number of integration events on the respective chromosomes, and is plotted on a different scale (in the Y-axis to the right) to the right.

2.7.2 Motif Detection

Increasing number of motif detection programs and algorithms have been designed and are available in the public domain. One of the most well known motif finding program is the `MEME` (Multiple Expectation-maximization for Motif Elicitation) (41). The 2kb length sequences obtained (described earlier on page 54), were subjected to the `MEME`, and following constraints were set viz.,

- Number of motifs to be found - 10
- Length of motifs to be found - minimum 5 to maximum 50
- Number of times a consensus sequence is expected to be in the dataset - total number of input sequences
- A consensus motif may be present zero or one or more than once per sequence
- Both the strands of the DNA sequence to be searched

All the parameters were decided by trial and error. The program *per se* is highly computation intensive, and it generates a *Regular Expression* for the motif from the input sequences. From the BH-series sequences (1) we obtained a set of ten motifs, the motifs and their sequences are shown in table 2.1.

Motif	Motif Sequence ($5' - 3'$)
1.	GGCGCGCGCCTGTAATCCCAGCACCTGCGGAGGCGCGAGGCGGGGGGGGATCA
2.	CCCCGGGTGGCGGGGATTGCAGGGATCTGCGATCACGCCAAGC
3.	CCAGCCTGGGCAACAACAGAGTGAGACCCCGTCT
4.	AGACGGGGTTTCACCATGTTGGCCAGGCTGG
5.	AAAAAAAAAAAAAAAATTAGCCGGGCCGTGGT
6.	CCCGGGCTCAAGTGATCCTCCCGCCTCAGCC
7.	CAGGCGTGAGCCACCACGCCCGGCTAATTTT
8.	CACGCCTGTAATCCCAGCTACTTGGGAGGCTGAGGCAGGAGGATCGCTTG
9.	TTTTTTTTTTTTTTTTTTTTTTGAGACGGAGTCTGCTCTGTC
10.	AGGTCAGGAGTTCGAG

Table 2.1: **Motifs obtained after processing the sequences through `MEME`:** The 2 kb flanking regions obtained (as described in the text) were used in the `MEME` to obtained motifs. Most of the default settings for the program were preserved. Ten most significant motifs were obtained under the default background model of the algorithm.

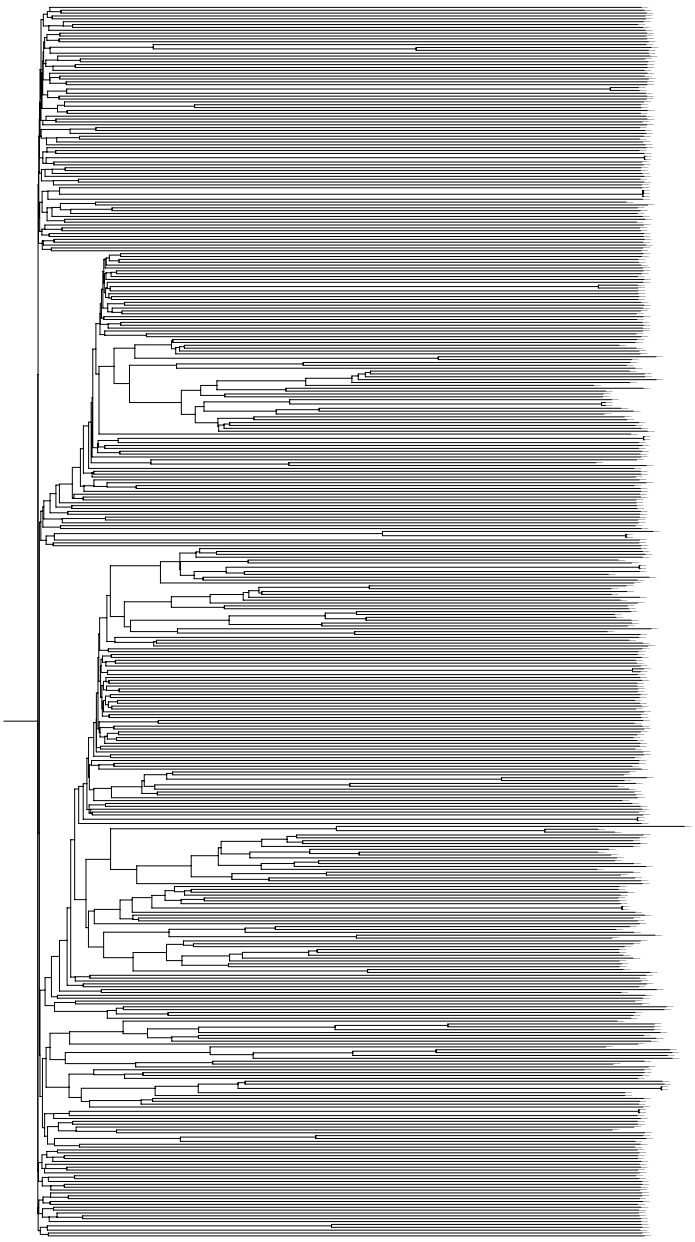


Figure 2.4: **Unrooted Tree obtained by Multiple Alignment of the Integration Sequences:** As described in the text 2000 bp flanking the invivo integration sites were downloaded and subjected to MSA using the clustal algorithm.

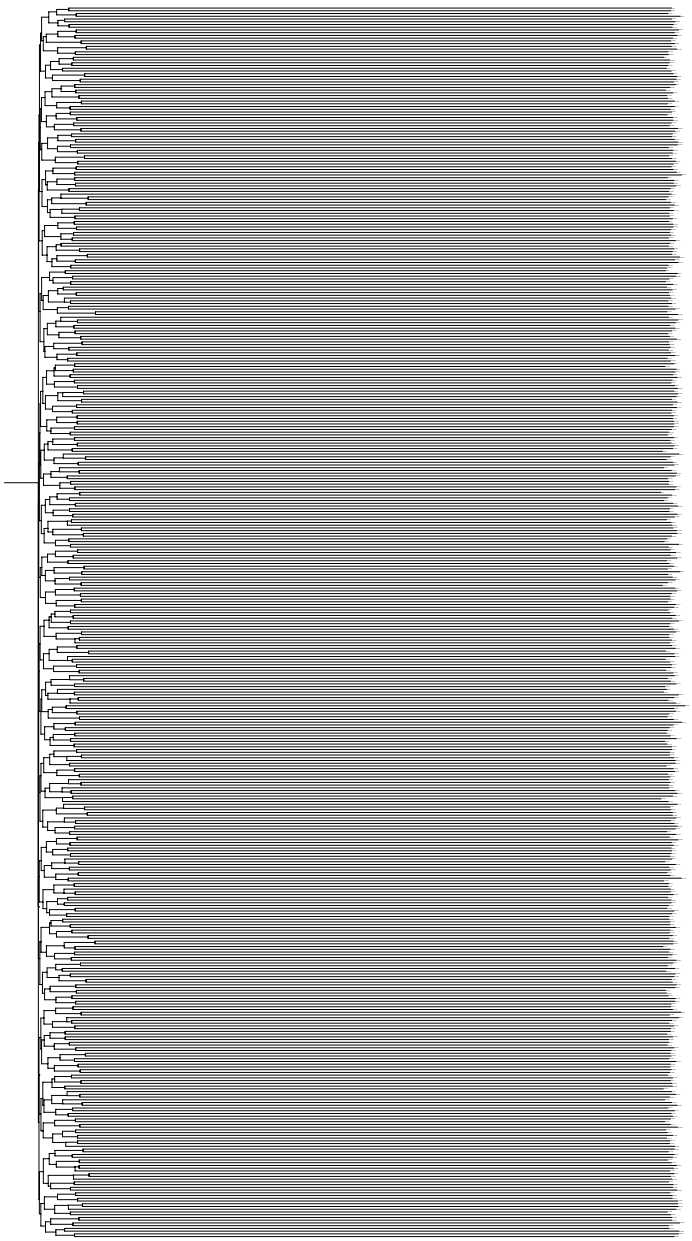


Figure 2.5: **Unrooted Tree obtained by Multiple Alignment random simulated sequences with human bias:** Briefly, 5^{th} order Markov Model obtained a human contig was used as described at the RSAT. 450 sequences each of length 2000 bp were generated and were used in a MSA exercise. The obtained distance matrix is visualized here as a unrooted tree.

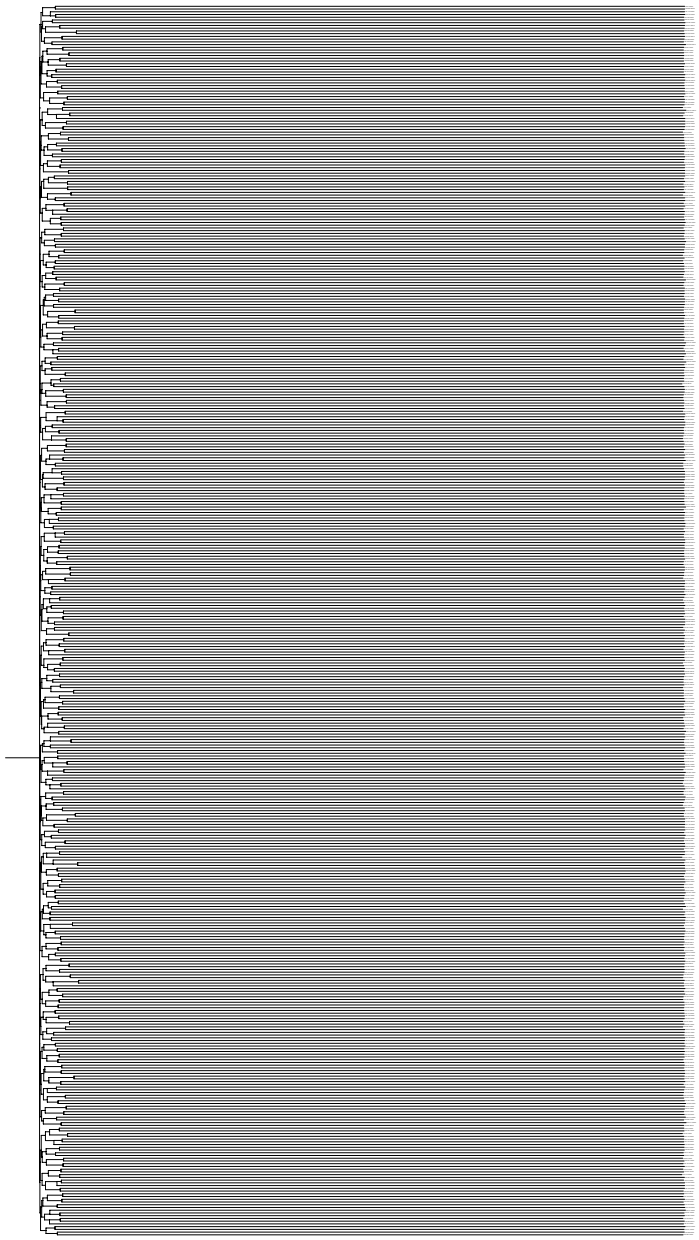


Figure 2.6: **Unrooted Tree obtained by Multiple alignment of random sequences of equal length:** Briefly, 450 sequences each of length 2000 bp were generated assuming equal probabilities for all the nucleotides. This figure shows a unrooted tree obtained after MSA of such sequences.

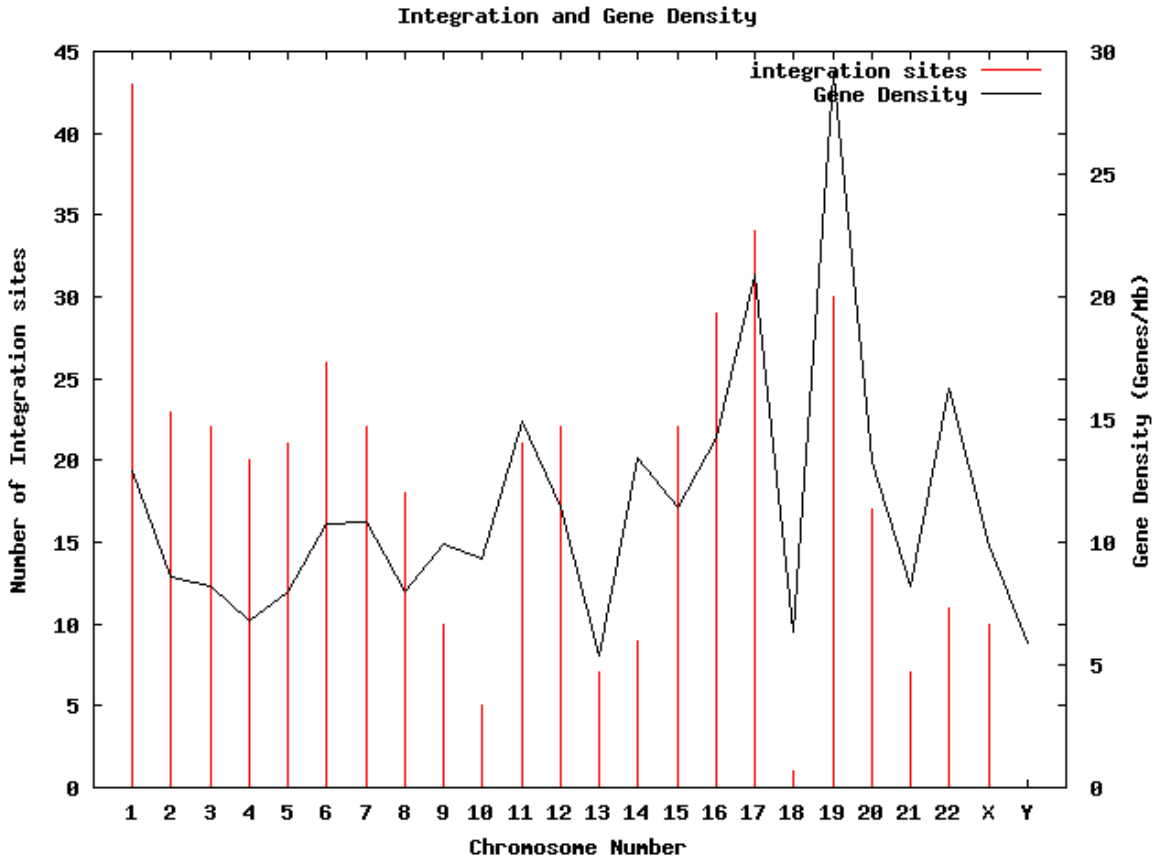


Figure 2.7: **Integrations sites and gene density are positively correlated:** In the Figure, the 'red' spikes denote the number of integration sites in a chromosome (the abscissa), plotted on the Y-axis on the left handside. The 'black' line denotes the gene density/Mb on each chromosome plotted on the Y-axis on the right handside for the same abscissa as before. It can be seen that the Gene density and number of integration sites are positively correlated.

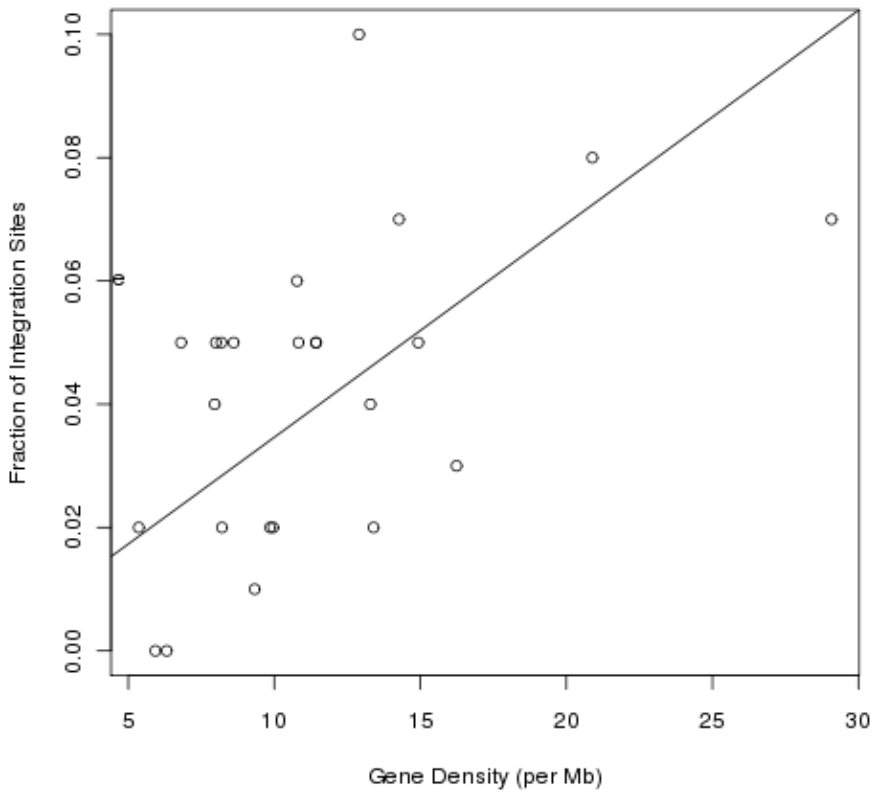


Figure 2.8: **Correlation between Gene density and integration sites:** In this figure the number of integration sites are plotted against the gene density. The line shows the linear positive correlation between the number integration sites and gene density.

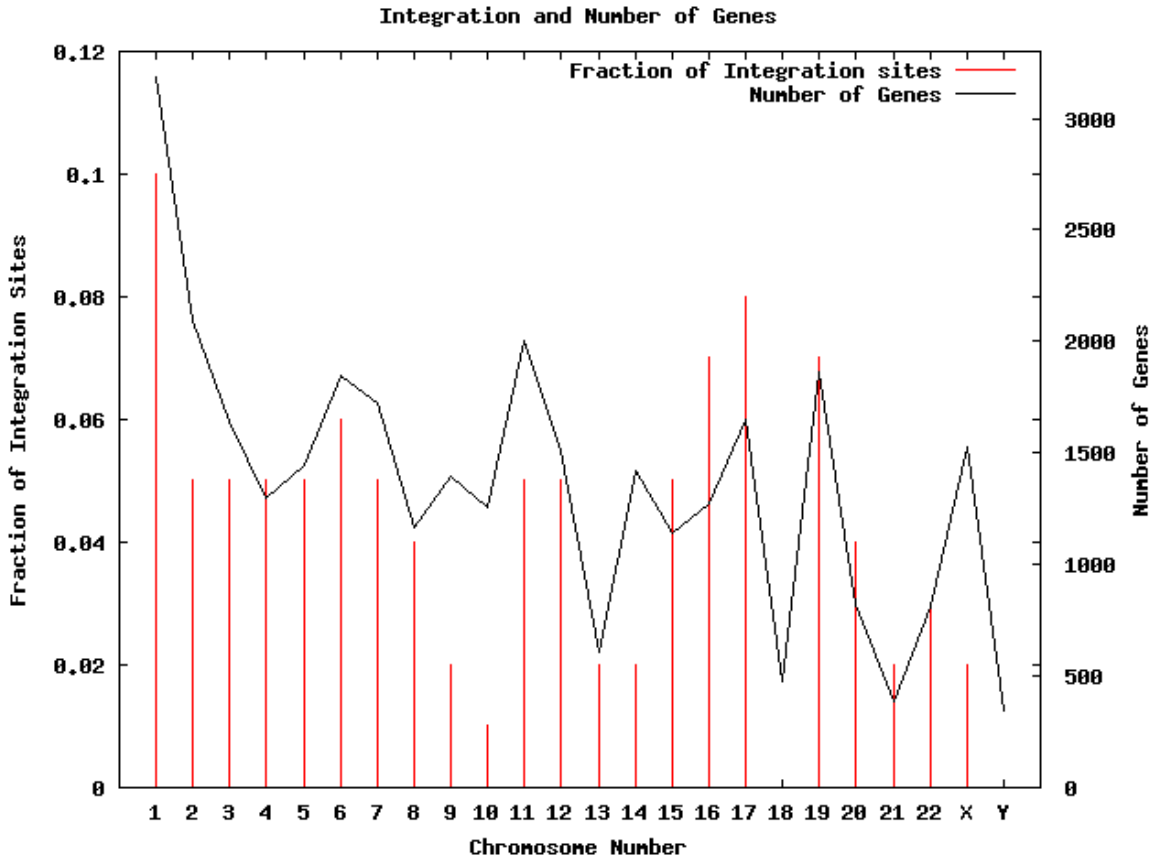


Figure 2.9: **Number of integration sites on a chromosome are directly proportional to number of genes:**The abscissa is the chromosome. The red impulses denote the fraction of integration sites present on each chromosome (plotted on Y-axis on the left hand side). The fraction is basically ratio of number of integration sites mapped to a chromosome to the total number of mapped integration sites. The black line (plotted on the Y-axis on the right hand side) shows number of genes present on each chromosome.

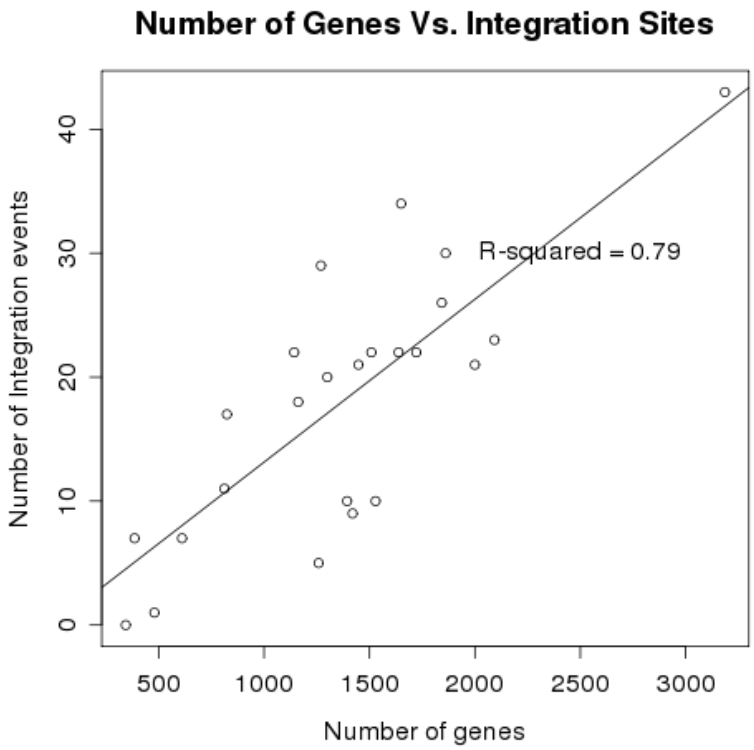


Figure 2.10: **Number of integration sites and number of genes are very highly positively correlated:** In this figure the number of genes is plotted on the abscissa and the fraction of integration sites associated with these number of genes a plotted on the ordinate. The positive correlation can be seen between the number genes and the integration sites.

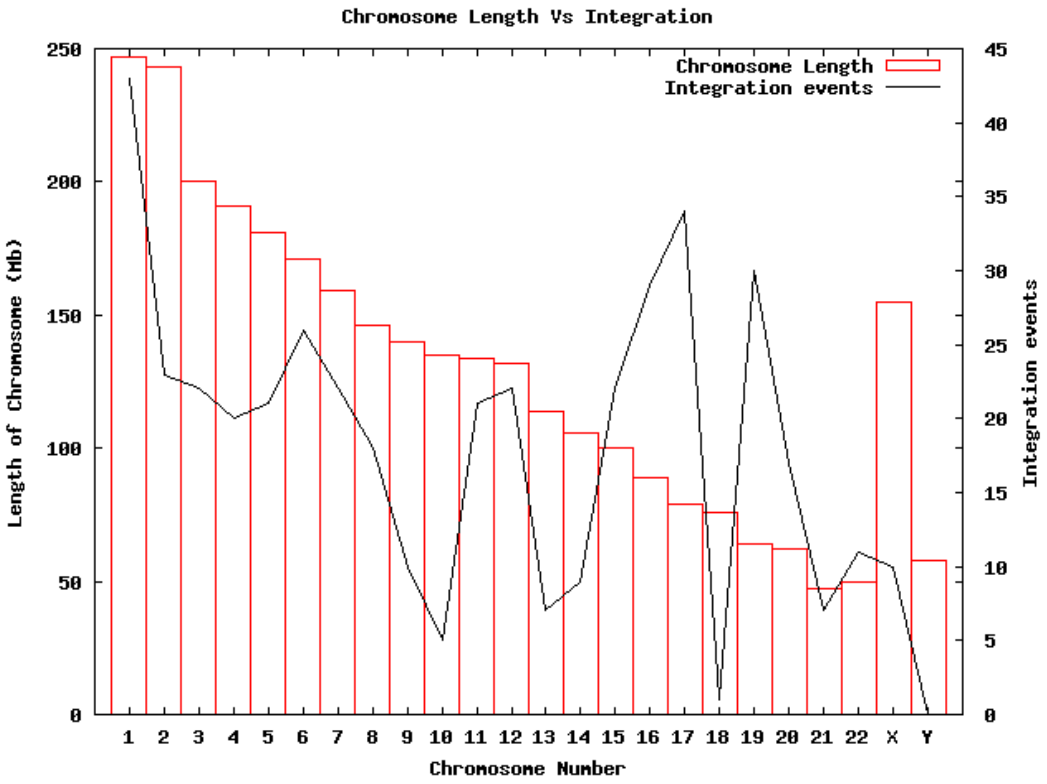


Figure 2.11: **Number of integration sites on a chromosome and its length are weakly correlated**: The red impulses denote the length of the chromosome and the black line denotes the number of integration sites present on the chromosome.

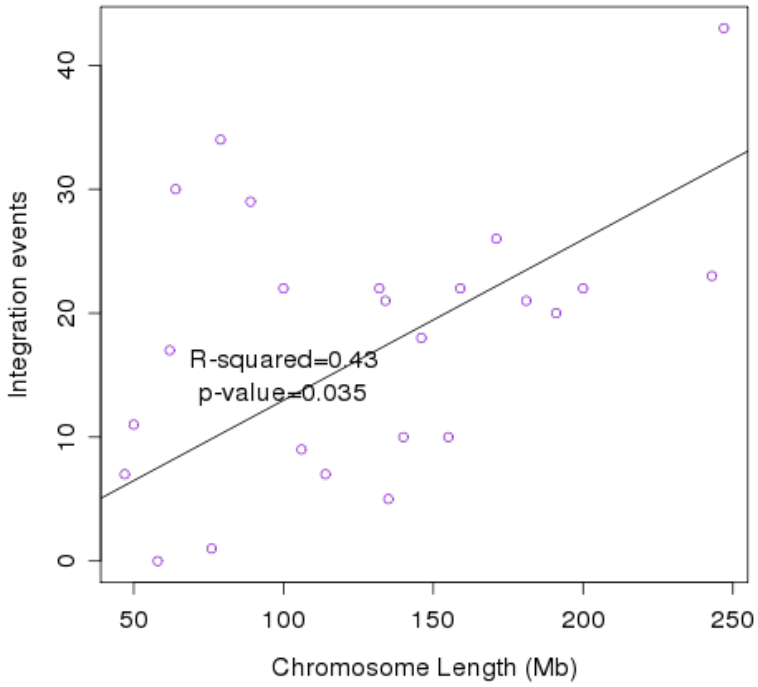


Figure 2.12: **Chromosome length and integration sites (correlation)**: The number of integration sites and length of chromosome are weakly positively correlated. In this figure the length of the chromosomes is plotted on the abscissa and the number of integration sites associated with these lengths of chromosmes are plotted on the ordinate. The positive correlation can be seen between the length of the chromosomes and the integration sites.

Motif No.	Motif Width	No. of Sites	Cumulative E-value	Average Occurrence
1	50	500	6.3×10^{-5907}	1.17
2	43	500	6.7×10^{-3529}	1.17
3	31	461	1.6×10^{-3413}	1.07
4	31	329	4.1×10^{-2511}	0.77
5	31	491	3.5×10^{-1608}	1.4
6	31	285	1.1×10^{-1539}	0.66
7	31	286	7.4×10^{-1491}	0.67
8	50	175	5.5×10^{-2015}	0.41
9	43	168	8.5×10^{-1516}	0.39
10	16	273	3.0×10^{-709}	0.64

Table 2.2: Statistics for the Motifs found in the BH series sequences.

Motif No.	Motif Width	No. of Sites	Cumulative E-value	Average Occurrence
1.	12	8	6.3e+009	0.02
2.	15	11	1.4e+010	0.02
3.	12	14	1.6e+010	0.03
4.	15	2	3.1e+010	0.004
5.	15	2	3.1e+010	0.004
6.	15	8	4.8e+009	0.02
7.	12	8	3.3e+010	0.02
8.	12	2	3.9e+010	0.02
9.	12	2	4.0e+010	0.004
10.	12	2	4.0e+010	0.004

Table 2.3: Statistics for the motifs detected in the control sequences.

It can be seen from Table 2.2[3], that there are 10 motifs, that show a very low E-value. E-value denotes probability that a given result is false positive, e.g. 6.3×10^{-5907}, means there is a probability of getting one false positive if 10^{5907} random sequences are analyzed. The sequences of these motifs were highly similar to the *Alu* repeats (See Section 2.8). When these sequences were passed through the `RepeatMasker` (42, and references therein). It was seen that many Alu repeats and other simple repeats were masked, after processing through the program. However, the motifs were still seen in the 'masked' sequences when used with the `MAST` (41).

As a control data set, use also used randomly generated/simulated sequences, which were human biased. These sequences were of identical length as that of the test sequences (obtained from (1)). The `MEME` (41) was used on these control sequences with same parameters that generated Table 2.2. The results obtained are summarized in Table 2.3. It can be seen when comparing the column 'Cumulative E-value' (for description see page 2.7.2) for Tables 2.2 and 2.3, that the E-value for the motifs detected in the control sequences is very high. In fact, it should be noted that the powers are all positive which means for each true positive detected there will be 10^n false positive results, where n is the E-value. Thus we can safely say that the motifs detected are of not statistically significant. Also the motifs are not seen in many sequences. Thus it can be said that the retroviral integration sites contain specific motifs, which are not seen in random sequences. Furthermore, it can be seen that irrespective of the sequence of the motif, occurrence in random sequences is rare. Whereas, in the invivo sequences the motifs are not only present as significant sequences, but they are also present in large more number of sequences.

2.8 *Alu*-like motifs are enriched in sequences flanking the reported HIV-1 integration sequences

As seen from the motif detection exercise, it is clear that the motifs detected are part of Alu repeats. We evaluated the role of the repetitive DNA in the selection of the integration site by the retroviral PIC. We used the sequences available in the public domain and reported

[3]The average occurrence denotes total number of occurrences of a given motif divided by total number of sequences taken for analysis

Sr. No.	Repeat Category	Number
1.	Present in BH series only	159
2.	Present in the ASLV integration sites only	29
3.	Present only in the PBMCs	10
4.	Present in BH series and the ASLV integration sequences	51
5.	Present in BH series and PBMCs	32
6.	Present in ASLV and PBMCs	4
7.	Common repeats present in all the data set	54
	Total	339

Table 2.4: Number of repeats present in different data sets and their combinations.

earlier by Schröder *et al.* and Mitchell *et al.* (1, 2). The data sets used and the preliminary statistics generated is given in Table 2.5.

These sequences were passed through the `CENSOR` program (43). The output of the program is in three parts, it returns list of repetitive elements present in each sequence and the position(s) in the sequence where the repetitive element is present. It can be seen from Table 2.4, the fraction of number of sequences containing repeat sequences is similar across the data sets. However, there is a difference in the number of repetitive sequences present exclusively to one set of sequences. Of these repetitive sequences, most prominent are the specific families of *Alu* sequences.

This prompted us to investigate further the possible role of Alu repeats in retroviral integration.

Series	Number of Sequences in the series	Number of sequences with repeats	Number of Repeats present	Average occurrence of Repeats per sequence	Maximum number of Repeats per sequence	Number of sequences without repeat(s)
BH Series	450	409	298	3.97	10	41
ASLV Integration sites	100	97	140	3.74	8	3
HIV integration sites in PBMCs	100	95	101	3.17	7	5

Table 2.5: **Summary of the occurrence of the Repeat DNA sequences in different data sets**: As described in the text, each sequence from each data set was processed using the `CENSOR` program. The outputs generated were parsed using scripts to determine sequence-repeat statistics and the results are summarized in this table.

2.8.1 More on Alu repeats

Less than 2% of the human genome 'codes' for something. Of the rest of the bulk of DNA the highly repetitive DNA sequences account for more than 50%. *Alu* elements are each a dimer of similar but not identical fragments of total size about 300 bp (44). Each element contains a bipartite promoter for RNA Polymerase III, a poly(A) tract located between the monomers, a 3′-terminal poly(A) tract, numerous CpG islands and is flanked by short direct repeats (45). *Alu* comprise more than 10% of the human genome and are capable of retroposition (45, 46). Insertion of an *Alu* element into a functionally important genome region or other *Alu*-dependent alterations of gene functions cause various hereditary disorders and are probably associated with carcinogenesis (45, 47, 48, 49). In total, 14 *Alu* families differing in diagnostic mutations are known (50). Some of these repeats and repeat families present in the human genome, are polymorphic and relatively recently inserted into new loci (51). *Alu* copies transposed during ethnic divergence of the human population are useful markers for evolutionary genetic studies (45, 47). *Alu* repeats have been associated with microsatellite repeats (44). It has been suggested that Alu repeats may be contributing to the origin of the microsatellite repeats (52). There are evidence to suggest that there is indeed a mechanism which enables the retroviruses and LTR-transposons to target specific chromosomal regions for integration (25). *Alu* has been deemed to be associated with retroviral integration (44). This has also helped make quantitative assays to estimate HIV integration (53). Another report (54) also suggests that the *Alu* repeats are associated with retroviral integration sites/ junctions and can be used to quantify retroviral integration.

2.9 Oligonucleotide analysis

In addition to the motif detection we also carried out analysis of the specific oligomers present in the integration sequences. This is a computationally less demanding method than motif detection. However it is limited to finding hexamers from a given set of sequences which are consistent and rare. The analysis was carried out as follows.,

1. Hexamers were identified in each sequence
2. Scores for each of the hexamers in the sequences were obtained using the Karlin's

approach (55), as follows:

(a) Dinucleotide probability distribution for a sequence was obtained i.e. $(p_{AA}, p_{AC}, \ldots, p_{TT})$

(b) Similarly probability of occurrence of each hexamer in a given sequence was obtained. For example, $p = p(ATTGAC) = p_{AC}.p_{GA}.p_{TG}.p_{TT}.p_{AT}$
Likewise probability for each hexamer was calculated based on its di-nucleotide composition.

(c) The probability was obtained under the assumption of randomness,
$q = p(ATTGAC) = (0.25)^6$.

(d) The score for a pattern was obtained using $s_i = log(q_i/p_i)$. Likewise, score for each hexamer was generated.

3. Frequency distribution of the scores was obtained. The data were scaled so that $s_i > 0$ for all is without the skewness of the distribution which is positive in majority of the sequences.

4. Weibull probability distribution fitted will to such data scores. The probability distribution function has two parameters p (shape) and q (scale) and is given by

$$f(s, p, q) = \frac{p}{q}\left(\frac{s}{q}\right)^{p-1} e^{(s/q)^p} \tag{2.1}$$

5. The parameters of the distribution were obtained for the score data of each sequence using Maximum Likelihood Estimation (MLE) method.

6. A threshold score s_0 was obtained by solving,

$$P(S \geq s_0) = e^{-(s_0/q)^p} = 0.05 \tag{2.2}$$

7. A subset of patterns for scores $S^{'} = \{s_i\}$ for $\{s_i\} \geq s_0$ was obtained

8. The patterns corresponding to these scores were identified in the sequences as 'rare' patterns.

9. The above steps from 3-8 were repeated for all the sequences in each group, thereby generating a list of patterns whose chances are rare in the respective sequences.

Such patterns were identified from each sequence of in vivo set (2kb) and those having more than 40% occurrence in the sequences of this set which are listed in the Table 2.6. The percentage occurrence of these patterns in sequences of original BH series are shown in column (II), while the occurrence in invitro sequences is shown in column (III) of Table 2.6. 450 sequences were randomly simulated using first order Markov chain (56) simulator and the occurrence of these patterns was observed in the sequences, which is as shown in column (IV) of Table 2.6. It can be seen that certain hexamers are rare, but consistently present in the BH series (1) of sequences.

S.No.	Pattern	In vivo		Invivo	Simulated**
		I	II	III	IV
1.	TCACGn (p1)	55.47	12.82	5.50	4.76
2.	CGTGAn (p2)	53.38	16.08	7.23	5.33
3.	GCGTGn (p3)	53.14	17.94	10.09	5.33
4.	CGAGAn (p4)	49.88	15.61	8.25	9.76
5.	TCTCGn (p5)	48.71	16.08	8.25	12.38
6.	CGCCTn (p6)	46.85	22.37	2.75	12.61
7.	CACGTn (p7)	45.22	11.65	5.50	5.00
8.	ACGTGn (p8)	44.75	9.09	3.66	6.90
9.	AGGCGn (p9)	44.05	19.34	7.33	6.90
10.	CGTGGn (p10)	43.12	12.82	6.42	5.47
11.	CTCGGn (p11)	41.72	15.15	4.58	6.67
12.	CACGCn (p12)	41.25	17.01	2.75	14.04
13.	GGCGCn (p13)	39.86	14.21	4.58	14.28
14.	CGTCTn (p14)	39.62	13.51	5.50	6.90

Table 2.6: Hexamer Analysis.

The occurrence of these patterns in the original BH (1) series sequences was observed. The table indicates the percentage of sequences possessing the above pattern at least once. The '*' indicate that the difference between the percentage of occurrence of all fourteen patterns of group I and IV, Table 2.6 is statistically significant with $p < 0.0001$ using $z - test$ for proportions. In other words, these patterns do occur with high proportions in in vivo sequences but in random sequences, their proportion was significantly low. It can be seen in Table 2.6, that the pattern has 'n' as the end nucleotide for each hexamer. We did a further analysis to check if there is any particular nucleotide that is preferred at the last position. The results are shown in Table 2.7.

Serial Number	**Pattern**	**Distribution of bases at n**				**Total Out of 429 sequences**
		A	C	G	T	
1.	TCACGn	64	92	31	51	238
2.	CGTGAn	37	28	79	87	229
3.	GCGTGn	81	28	77	42	228
4.	CGAGAn	22	89	27	76	214
5.	TCTCGn	69	48	60	32	209
6.	CGCCTn	29	71	80	21	201
7.	CACGTn	40	41	65	48	194
8.	ACGTGn	57	45	41	49	192
9.	AGGCGn	13	49	45	82	189
10.	CGTGGn	34	44	21	86	185
11.	CTCGGn	9	87	58	25	179
12.	CACGCn	31	103	12	31	177
13.	GGCGCn	46	63	45	17	171
14.	CGTCTn	28	84	30	28	170

Table 2.7: Probability chart showing probability of finding a particular nucleotide at the last position on the hexanucleotide.

It can be further seen that the position number 6 in the hexamers, is fuzzy to some extent, however, there is still a base preference at that position in each of the patterns mentioned in Table 2.7.

The sequences obtained from the Schroeder *et al.* (1), were classified using Kmeans clustering method. In every class some patterns indicate dominant occurrence. The algorithm has to be provided with present number of classes (in a sense it is a guided clustering). In this case the sequences were clustered into 15 classes. The assumption being, ones dominant pattern per class, and one class where none are dominant or all are equally dominant. Thus it can be concluded that the sequences preferred by the HIV provirus for the integration are enriched in certain oligomers.

2.10 Pattern recognition in retro-viral integration genomic sequences

We initially started with AY-series (30), that contains the HIV integration sites as occurring in HeLa cell line and H9 cell line. There is another set deposited by the same group mentioned in the same paper that is the data for the integration sites of MLV in HeLa cell line. Our initial data was taken only from the those cell lines where 'HIV' integrations were studied.

2.10.1 Methodology

We started with looking for the 'consistent' tetramers in the given dataset. There are about 519 such sequences. However, the variation in their length is very high (length from 7 bases to 250+ bases). For the ease of handling of data and in view of further interpretation of the results, we decided to utilize only those sequences with length $\geq$ 100 bases. We used 230 such sequences for our analysis. In these sequences we counted the number of occurrences of each tetramer (with alphabet size of 4 i.e. A, C, G and T, there are 4^4 i.e. 256 possible tetramers), that occurs in each of the sequences. We found only four such tetramers. So we decided to look for the trimers ($4^3 = 64$ possibilities). We found that 16 trimers were present in more than 80% of the sequences. When we tried to look for the same trimers in the omitted sequences (sequences less than 100 bases in length), the same trimers were

found to be consistently present in most of the sequences. We further decided to study these particular 16 trimers. Most of these trimers are AT-rich. Multiple sequence alignment on these sequences did not yield interpretable results. Thus, we decided to embark upon a different and somewhat novel approach.

We substituted all the consistent trimers with single lettered code such that we now had a set of sequences which were represented only in terms of the consistent 3-mers present in the dataset. We used single letter codes for the amino acids to represent these trimers. We performed a multiple alignment of such substituted sequences. At this point many patterns appeared from the alignment. In particular only five trimers could be seen which were present consistently in most of the sequences, were at a similar position in the different sequences and were separated by a similar distance from each other across the sequences. The following five trimers were found to be similar in distribution across the sequences viz. C/GAA, AAA, ATT, TTT. Of these we found 2 sets of 2 trimers each that always occurred next to each other. These were C/GAA & AAA and ATT & TTT. So we took them to be C/GAAA and ATTT. There was also a AAA occurring away from both the tetramers in the middle. With this information we constructed the following regular expression `[CG]AAA.*AAA.*ATTT`. Everything within the square brackets '[ ]' is an alteration. This regular expression will match everything that starts with a C or a G and is followed by three As, followed by any nucleotide for any number of times, followed by three As followed by any nucleotide for any number of times, followed by the sequence ATTT. This is a predominantly AT-rich sequence. However, because of the '`.*`' construct in the regular expression, it is 'greedy' and will match with the largest length possible for any given query sequence. To make this regular expression stringent, we decided to include a distance parameter instead of the '`.*`'. So we replace '`.*`' with '`.{lowerlimit,upperlimit}`' where the lower limit is the distance in terms of number of characters that are present between two given sub-sequences. It was decided to look at the distance profile of the distance between the two motifs at the end and the AAA in the middle. A script was written to compute all possible distances between these two in all the sequences. We looked at the bi-variate distribution pattern of these sequences. The bivariate distribution showed a positive skew, and a peak was observed at the distance of 0 - 50 for both the regions. As the length of first motif is 4 bases and that of second is 3 bases, we took

the distances between 5 to 50 instead of 0 to 50. We also found 2 more peaks at 51 - 100 and 0 - 50 respectively, for first distance the second distances. So our Regular Expression now became:

```
[CG]AAA.{5,50}AAA.{5,50}ATTT
```

With the distance between the sequences thus defined, we decided to look for consistent motifs in the 5 - 50 base long region that separates the given three subsequences. To obtain those sequences specifically we changed the regular expression to

```
[CG]AAA(.{5,50})AAA(.{5,50})ATTT
```

According to the regular expression syntax, the region enclosed by round brackets can be isolated/consumed. So we gathered such 'spacer' sequences from the various data sets we had downloaded from the NCBI (1, 2, 30). We call the 'spacer' sequence that separates [CG]AAA from AAA as the left hand sequence and the one between the AAA and ATTT as the right hand sequence.

Validation We looked at the 'consistent' trimers present in the in between sequences. There was a clear pattern that was seen in the BH series sequences (1) and the AY (2) series sequences. In the left hand sequence an 'AAA' was clearly seen to be present in majority of the sequences. On the right hand sequence a 'TTT' was seen to be present in the majority of the sequences. We wrote a script that will consider a true match if and only if the left hand side contained 'AAA' and right hand side contained 'TTT'. The results obtained clearly showed that the BH series is different than the randomly generated sequences. However, statistically only the left hand side gave us unambiguous results, there was a lot of ambiguity seen in the right hand side sequence. Data generated only from the left hand side was taken for further analysis.

2.10.2 Analysis using regular expression

We ranked the occurrences of the consistent trimers in the left hand side sequences. We did Spearmanns rank correlation test. We plotted a classification tree taking all the 6 datasets viz.,

1. HIV with SupT1 (the BH series by Schroder *et al.* (1))
2. HIV with PBMCs (Mitchell *et al.* (2))
3. HIV with HeLa and H9 (the AY series, Wu *et al.* (30))
4. HIV with IMR90 (PLOS Biology paper, (2))
5. MLV with HeLa (Wu *et al.* (30))
6. ASLV with 293T-TVA ((2))

The last 2 datasets i.e. the MLV on human cell line and ASLV on 293T-TVA cell line were clearly differentiated from the rest of the sequences. AY - aeries was shown to be closer to the BH - series, IMR90 was close to the PBMCs. Thus we have dataset wherein three kinds of retroviruses were used to infect six different kinds of cell lines. The Spearman rank is calculated as follows:

$$Rs = 1 - \left(\frac{6\sigma d^2}{n^3 - n} \right) \tag{2.3}$$

Where, Rs is the Spearman rank, d is the difference between the ranks, n is the number of ranks in the data. We found that the the types of n-mers enriched in the flanking regions of the insertion sites were distinct for different types of viruses. This again points to recognition of some kind of higher-order structure in the genome rather than a specific sequence. We found no particular pattern/distribution for any of the specifically enriched 3-mers.

2.11 Conclusion and Summary

Thus it can be concluded that the HIV integration target sites are specific within the genome. Moreover, similar viruses infecting similar hosts target similar sites. Whereas pseudotyped viruses target atypical sequences (which are usually not encountered in the natural host). Given that the sequences flanking the HIV integration sites have specific properties, it should be possible to computationally predict integration sites using neural-network like methodologies.

References

[1] A. R. Schroder, P. Shinn, H. Chen, C. Berry, J. R. Ecker, and F. Bushman. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, 110(4):521–9, 2002.

[2] R. S. Mitchell, B. F. Beitzel, A. R. Schroder, P. Shinn, H. Chen, C. C. Berry, J. R. Ecker, and F. D. Bushman. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol*, 2(8):E234, 2004.

[3] B. Irwin, M. Aye, P. Baldi, N. Beliakova-Bethell, H. Cheng, Y. Dou, W. Liou, and S. Sandmeyer. Retroviruses and yeast retrotransposons use overlapping sets of host genes. *Genome Res*, 15(5):641–54, 2005.

[4] A. Rambaut, D. Posada, K. A. Crandall, and E. C. Holmes. The causes and consequences of HIV evolution. *Nat Rev Genet*, 5(1):52–61, 2004.

[5] X. Zhang, Y. Hakata, Y. Tanaka, and H. Shida. CRM1, an RNA transporter, is a major species-specific restriction factor of human T cell leukemia virus type 1 (HTLV-1) in rat cells. *Microbes Infect*, 8(3):851–9, 2006.

[6] T. Ikeda, J. Shibata, K. Yoshimura, A. Koito, and S. Matsushita. Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. *J Infect Dis*, 195(5):716–25, 2007.

[7] S. Bour, R. Geleziunas, and M. A. Wainberg. The human immunodeficiency virus type 1 (HIV-1) CD4 receptor and its central role in promotion of HIV-1 infection. *Microbiol Rev*, 59(1):63–93, 1995.

[8] J. Lehmann-Che and A. Saib. Early stages of HIV replication: how to hijack cellular functions for a successful infection. *AIDS Rev*, 6(4):199–207, 2004.

[9] R. Belshaw, A. L. Dawson, J. Woolven-Allen, J. Redding, A. Burt, and M. Tristem. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. *J Virol*, 79(19):12507–14, 2005.

[10] B. Bowerman, P. O. Brown, J. M. Bishop, and H. E. Varmus. A nucleoprotein complex mediates the integration of retroviral DNA. *Genes Dev*, 3(4):469–78, 1989.

[11] B. A. Castro, C. Cheng-Mayer, L. A. Evans, and J. A. Levy. HIV heterogeneity and viral pathogenesis. *AIDS*, 2 Suppl 1:S17–27, 1988.

[12] M. Alfano, A. Crotti, E. Vicenzi, and G. Poli. New players in cytokine control of HIV infection. *Curr HIV/AIDS Rep*, 5(1):27–32, 2008.

[13] R. A. Katz, J. G. Greger, and A. M. Skalka. Effects of cell cycle status on early events in retroviral replication. *J Cell Biochem*, 94(5):880–9, 2005.

[14] E. H. Humphries and J. M. Coffin. Rate of virus-specific RNA synthesis in synchronized chicken embryo fibroblasts infected with avian leukosis virus. *J Virol*, 17(2):393–401, 1976.

[15] R. A. Katz, J. G. Greger, K. Darby, P. Boimel, G. F. Rall, and A. M. Skalka. Transduction of interphase cells by avian sarcoma virus. *J Virol*, 76(11):5422–34, 2002.

[16] W. J. Swiggard, C. Baytop, J. J. Yu, J. Dai, C. Li, R. Schretzenmair, T. Theodosopoulos, and U. O'Doherty. Human immunodeficiency virus type 1 can establish latent infection in resting CD4+ T cells in the absence of activating stimuli. *J Virol*, 79(22):14179–88, 2005.

[17] D. N. Vatakis, G. Bristol, T. A. Wilkinson, S. A. Chow, and J. A. Zack. Immediate activation fails to rescue efficient human immunodeficiency virus replication in quiescent CD4+ T cells. *J Virol*, 81(7):3574–82, 2007.

[18] C. M. Farnet and F. D. Bushman. HIV-1 cDNA integration: requirement of HMG I(Y) protein for function of preintegration complexes in vitro. *Cell*, 88(4):483–92, 1997.

[19] E. Agbottah, L. Deng, L. O. Dannenberg, A. Pumfery, and F. Kashanchi. Effect of SWI/SNF chromatin remodeling complex on HIV-1 Tat activated transcription. *Retrovirology*, 3:48, 2006.

[20] R. G. Beiko and R. L. Charlebois. GANN: genetic algorithm neural networks for the detection of conserved combinations of features in DNA. *BMC Bioinformatics*, 6:36, 2005.

[21] A. Sloman and J. Chappell. Computational cognitive epigenetics. *Behav Brain Sci*, 30(4):375–6, 2007.

[22] M. K. Lewinski and F. D. Bushman. Retroviral DNA integration–mechanism and consequences. *Adv Genet*, 55:147–81, 2005.

[23] F. D. Bushman and R. Craigie. Sequence requirements for integration of Moloney murine leukemia virus DNA in vitro. *J Virol*, 64(11):5645–8, 1990.

[24] F. D. Bushman, T. Fujiwara, and R. Craigie. Retroviral DNA integration directed by HIV integration protein in vitro. *Science*, 249(4976):1555–8, 1990.

[25] F. D. Bushman. Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. *Cell*, 115(2):135–8, 2003.

[26] K. Doi, X. Wu, Y. Taniguchi, J. Yasunaga, Y. Satou, A. Okayama, K. Nosaka, and M. Matsuoka. Preferential selection of human T-cell leukemia virus type I provirus integration sites in leukemic versus carrier states. *Blood*, 106(3):1048–53, 2005.

[27] T. Mailund, S. Besenbacher, and M. H. Schierup. Whole genome association mapping by incompatibilities and local perfect phylogenies. *BMC Bioinformatics*, 7:454, 2006.

[28] G. D. Trobridge, D. G. Miller, M. A. Jacobs, J. M. Allen, H. P. Kiem, R. Kaul, and D. W. Russell. Foamy virus vector integration sites in normal human cells. *Proc Natl Acad Sci U S A*, 103(5):1498–503, 2006.

[29] M. K. Lewinski, M. Yamashita, M. Emerman, A. Ciuffi, H. Marshall, G. Crawford, F. Collins, P. Shinn, J. Leipzig, S. Hannenhalli, C. C. Berry, J. R. Ecker, and F. D. Bushman. Retroviral DNA integration: viral and cellular determinants of target-site selection. *PLoS Pathog*, 2(6):e60, 2006.

[30] X. Wu, Y. Li, B. Crise, and S. M. Burgess. Transcription start regions in the human genome are favored targets for MLV integration. *Science*, 300(5626):1749–51, 2003.

[31] H. Liu, E. C. Dow, R. Arora, J. T. Kimata, L. M. Bull, R. C. Arduino, and A. P. Rice. Integration of human immunodeficiency virus type 1 in untreated infection occurs preferentially within genes. *J Virol*, 80(15):7765–8, 2006.

[32] S. Carteau, C. Hoffmann, and F. Bushman. Chromosome structure and human immunodeficiency virus type 1 cDNA integration: centromeric alphoid repeats are a disfavored target. *J Virol*, 72(5):4005–14, 1998.

[33] Y. C. Bor, F. D. Bushman, and L. E. Orgel. In vitro integration of human immunodeficiency virus type 1 cDNA into targets containing protein-induced bends. *Proc Natl Acad Sci U S A*, 92(22):10334–8, 1995.

[34] J. F. Hughes and J. M. Coffin. Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. *Genetics*, 171(3):1183–94, 2005.

[35] C. N. Johnson and L. S. Levy. Matrix attachment regions as targets for retroviral integration. *Virol J*, 2:68, 2005.

[36] I. Leclercq, F. Mortreux, A. S. Gabet, C. B. Jonsson, and E. Wattel. Basis of HTLV type 1 target site selection. *AIDS Res Hum Retroviruses*, 16(16):1653–9, 2000.

[37] T. J. Schuh, H. Ahrens, M. D. Mogensen, J. Gorski, and G. C. Mueller. Polyclonal antibodies from rabbits and chickens against the estrogen receptor and related peptides. Use in the affinity isolation of estrogen receptors and the retrieval of chromatin fragments associating with estrogen receptors. *Receptor*, 2(2):93–107, 1992.

[38] L. A. Dickinson, T. Joh, Y. Kohwi, and T. Kohwi-Shigematsu. A tissue-specific MAR/SAR DNA-binding protein with unusual binding site recognition. *Cell*, 70(4):631–45, 1992.

[39] W. R. Pearson. Effective protein sequence comparison. *Methods Enzymol*, 266:227–58, 1996.

[40] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990.

[41] T. L. Bailey, M. E. Baker, and C. P. Elkan. An artificial intelligence approach to motif discovery in protein sequences: application to steriod dehydrogenases. *J Steroid Biochem Mol Biol*, 62(1):29–44, 1997.

[42] J. A. Bedell, I. Korf, and W. Gish. MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics*, 16(11):1040–1, 2000.

[43] J. Jurka, V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, and J. Walichiewicz. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, 110(1-4):462–7, 2005.

[44] G. R. Daniels and P. L. Deininger. Integration site preferences of the Alu family and similar repetitive DNA sequences. *Nucleic Acids Res*, 13(24):8939–54, 1985.

[45] Khitrinskaia IIu, V. A. Stepanov, and V. P. Puzyrev. [Alu repeats in the human genome]. *Mol Biol (Mosk)*, 37(3):382–91, 2003.

[46] R. Cordaux, D. J. Hedges, S. W. Herke, and M. A. Batzer. Estimating the retrotransposition rate of human Alu elements. *Gene*, 373:134–7, 2006.

[47] E. I. Rogaev. Simple human DNA-repeats associated with genomic hypervariability, flanking the genomic retroposons and similar to retroviral sites. *Nucleic Acids Res*, 18(7):1879–85, 1990.

[48] J. P. Jones, 3rd, M. N. Kierlin, R. G. Coon, J. Perutka, A. M. Lambowitz, and B. A. Sullenger. Retargeting mobile group II introns to repair mutant genes. *Mol Ther*, 11(5):687–94, 2005.

[49] L. L. Chen, J. N. DeCerbo, and G. G. Carmichael. Alu element-mediated gene silencing. *EMBO J*, 27(12):1694–705, 2008.

[50] L. R. Simard, J. Viel, M. Lambert, G. Paradis, E. Levy, E. E. Delvin, and G. A. Mitchell. The Delta¿15 Kb deletion French Canadian founder mutation in familial hypercholesterolemia: rapid polymerase chain reaction-based diagnostic assay and prevalence in Quebec. *Clin Genet*, 65(3):202–8, 2004.

[51] V. V. Kapitonov, I. A. Shakhmuradov, and I. A. Kolchanov. [Evolution of Alu repeats. Imitation model]. *Genetika*, 25(6):1111–8, 1989.

[52] S. S. Arcot, Z. Wang, J. L. Weber, P. L. Deininger, and M. A. Batzer. Alu repeats: a source for the genesis of primate microsatellites. *Genomics*, 29(1):136–44, 1995.

[53] A. Brussel, O. Delelis, and P. Sonigo. Alu-LTR real-time nested PCR assay for quantifying integrated HIV-1 DNA. *Methods Mol Biol*, 304:139–54, 2005.

[54] O. Delelis, A. Brussel, and P. Sonigo. Quantification of HFV-integrated DNA in human cells by Alu-LTR real-time PCR. *Methods Mol Biol*, 304:155–70, 2005.

[55] S. Karlin. Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol*, 1(5):598–610, 1998.

[56] J. van Helden. Regulatory sequence analysis tools. *Nucleic Acids Res*, 31(13):3593–6, 2003.

Chapter 3

On Selecting Proper Control Sequences for Motif Detection Exercises

3.1 Introduction

A number of well-established probabilistic methodologies (1, and references therein), such as `MEME`, `SeSiMCMC`, `MotifSampler`, etc., attempt to find consistent and statistically significant patterns in a set of biological sequences (DNA, RNA, or protein). Such consistent patterns are often referred to as *motifs*. It is well-known that genomic motifs often represent highly conserved elements that may have important regulatory/functional roles. For example, genomic motifs could represent *cis* regulatory elements (2), retroviral insertion sites (3), etc. RNA motifs could represent pseudoknot structures that are now known to play important role in translation regulation, a feature that has been utilized in drug design (4). Motifs in proteins (seen as sequences of amino acids) often represent binding sites for interacting partners or chemically active sites – again a feature important for drug design (5). From an evolutionary perspective, motifs that are common across species are useful for deciphering phylogenetic relationships (6, 7).

Motif detection methodologies have become an important tool for biological sequence analysis primarily due to the size of the commonly available sequence data. Recent high-throughput technologies have also created a need for the automation of data analysis pipelines that include motif detection as an important step.

Position	A	C	G	T	Motif
1	0.000000	1.000000	0.000000	0.000000	C
2	0.000000	1.000000	0.000000	0.000000	C
3	0.000000	0.002381	0.000000	0.997619	T
4	0.000000	1.000000	0.000000	0.000000	C
5	0.626143	0.075790	0.191495	0.106572	A
6	0.000000	0.005125	0.994875	0.000000	G
7	0.000000	1.000000	0.000000	0.000000	C
8	0.000000	1.000000	0.000000	0.000000	C
9	0.000000	0.004610	0.000000	0.995390	T
10	0.000000	1.000000	0.000000	0.000000	C
11	0.000000	1.000000	0.000000	0.000000	C
12	0.000000	0.810820	0.000000	0.189180	C

Table 3.1: This table depicts a PSPM for the first motif as shown in Table 3.4. The general structure of the PSPM is as follows, it has number of rows equal to the length of the motif and number of columns equal to the length of the alphabet. As we are dealing with DNA sequences the alphabet size is 4 and hence there are 4 columns. The number in the first column denotes the position of the base. Each subsequent column shows the probability of occurrence of the particular base (column heading) at the particular position.

In general, probabilistic motif search methodologies attempt to infer the unknown locations of a motif in the user-supplied sequence data, by comparing the data with the *background* in a probabilistic manner. Simultaneously, they estimate parameters of the *motif model* (e.g. see Table 3.1).

If required, the same methodologies can additionally estimate parameters of the *background model*. Alternatively, a motif detection algorithm can be seen as a device for arriving at a partition of given input sequences into the *motif* part and the *background* part.

This is achieved by formulating the motif detection problem as a missing data problem (8, 9). In the context of motif detection, the input sequences in which motifs are to be discovered (and motif model parameters to be estimated) is the observed data. Missing observations, in this context, are the locations of one or more such motifs in the observed sequence data. The missing-data problem is formulated in terms of the likelihood function of the motif model parameters conditional upon the observed and the unobserved data. In a Bayesian formulation, the likelihood function gets replaced by the posterior distribution that incorporates any prior information about model parameters and unobserved data in addition to the likelihood function. To estimate model parameters and discover the most likely value of the motif locations, the likelihood function (or the posterior distribution) is either maximized, or sampled.

Deterministic maximization is usually performed using the *expectation-maximization* (EM) algorithm (10). Sampling-based approaches usually resort to a variant of the Markov Chain Monte Carlo (MCMC) method called Gibbs sampling (11). Detailed development of probabilistic motif detection algorithms of either type can be found in, e.g., (12, 13, 11, 14). A survey of some of the most prominent motif detection tools can be found in Section 3.1.2.

The work presented in this chapter deals with the choice of the background for probabilistic motif detection. In the rest of this section, we discuss some of the key concepts and issues related to probabilistic motif detection. Materials and methods are discussed in Section 3.3, and results are presented in Section 3.4. This chapter ends with a summary (Section 3.5) of the key results of this work and conclusions drawn from them.

3.1.1 Probabilistic Models of Genomic Sequences

Probabilistic models of genomic DNA sequences are motivated by the fact that real genomic DNA sequences are highly complex, hierarchically-organized entities that are shaped primarily through the forces of biological evolution. Furthermore, sequences of functionally distinct components of genomic DNA such as intergenic regions, promoters, genes, introns, exons, and structural elements (centromeres, telomeres) all have very different statistical properties. Finally, both short- and long-range correlations (15, 16) exist within the primary sequence of a genome for a variety of reasons, known or unknown. For instance, while short-range correlations on the length scale of say a gene or a promoter are usually associated with the complex machinery for gene expression and its control, one of the reasons for long-range correlations could be the chromatin context (i.e., hierarchical coiling of the primary DNA strand into highly compact higher-order structures) that can bring distant parts of the primary sequence in close physical proximity thus inducing functional correlations between them (17).

Perhaps the most useful models of genomic DNA sequences have to be probabilistic in nature given the complexity and variability of genomic DNA, as also the lack of precise knowledge about genomic DNA as a highly complex and hierarchical system. The simplest probabilistic model of a genomic DNA sequence is that which assumes independent, identically-distributed (IID) base alphabet (`A,C,G,T`) with pre-specified probabilities. However, the presence of long-range correlations in a genomic DNA sequence necessitates the

use of models that allow for greater complexity. The simplest class of models that allow for inclusion of sequential correlations in a systematic manner are the *Markov models* (18). Probabilistic models that allow for greater complexity to be modeled include hidden-Markov models and stochastic grammars (19, 18, see for an overview).

A Primer on Markov Models

The order-k Markov model of an unending (practically, very long) genomic sequence is defined as the set of probabilities for a single base `A, C, G` or `T` to follow an oligonucleotide sequence of length k. The complete specification of a Markov model of order k thus involves specification of 4^k (i.e., the number of possibilities for oligonucleotides of length k) $\times$ 4 (i.e., the number of bases) probabilities, and can be represented as a $4^k \times 4$ matrix. An order-0 Markov model thus corresponds to the aggregate probabilities of occurrence of the four DNA bases in the sequence they represent. Markov models of orders 0, 1, and 2 are illustrated for a randomly-picked human DNA sequence of length 2400 in Fig. 3.1.

In the language of probability theory, each entry in this matrix corresponds to the conditional probability $P(b|b_1b_2\ldots b_k)$ of base b to follow oligonucleotide $b_1b_2\ldots b_k$, where both b and $b_i, 1 \leq i \leq k$, assume values from the set of DNA alphabets {`A, C, G, T`}. These conditional probabilities are sometimes referred to as *transition probabilities*. By construction, each row of a Markov matrix/model is normalized; i.e., the sum of all entries of a row is 1. Parameters of an order-k Markov model of a genomic sequence (i.e., all entries of the matrix as described above) are estimated from sufficiently long DNA sequences by counting all occurrences where a specific oligonucleotide sequence $b_1b_2\ldots b_k$ precedes given base b.

Visualization of a Markov model

Jeffrey devised a method to represent genomic sequences in form of *Chaos Game Representation* (CGR), the construction of which is described in detail in (*Jeffery*, 90) (22), and illustrated in Fig. 3.2. A CGR plotted at a k-mer resolution is equivalent to a Markov model of the $k-1$ order (22, and references therein): The CGR is thus a useful way of visualizing an order-k Markov model.

A sequence of length 2400 picked from a random location in the human genome:

```
>ref|NT_006713.14|Hs5_6870:15287525-15289625 Homo sapiens chromosome 5 genomic contig, reference assembly
ATTTTCTTCTGCTTACTTTGAGTTTAATTAGCCCTCCTTTATCTAGTTACCTAAAATGGAAGCTTAGGTTTTTTAAGTCT
ATTTCCCTTTCTAATATATGCAACCAATGTTATAAATTTCCCTCTATGCACTGCTTTTGCTGCCTCTCATAAATTTTGGT
AAGTTATGTTTTCATTTCATTTAGTTCAAAGTATTTTTAATTTCTCTTCAGATTTTTCTTTTGACCCATGTGTTATTTAG
AAGTATGTTGTTTATGGCCAGGTGTGGTGGCTCACACCTGTAATCCCAGGATTTTGGTAGGCCCAGGAGACCAGATTGCT
TGAGCTCAGGAGTTCAAGGCCAGCCTGTGCAACATGGAGAAACCTCGACTCTACAAAAAAAAGAAAAATCAGCCAGGTGT
GGTGGCACATGCTTGTAGTCCCAGCTACTTAGGAGGTTGAGGTGGGAGGATTGCTTGAGCCCAGGAAGCAGAGGATGCAG
TGAGCTATGATTTTGCCACTACACTCCAGCCTCAGTGACAGAGTAAGACCCTGTCTTAAAAAAAAAAAAAAAAAAAAGT
ATGTTGCTTAATCTCCACATATTGTTTGGGGCTTTCCAGTTATCTTGCTGTGATTGATTTCTAGATTAATTCCATGTGGT
TTGAGAACAGATATTGTATGATATCTATTCTTTTAAATTTGTTAAGATGTGTTTAATGGCCCAGAATGTGGTCTGTCTTG
GTAGCTGTTCCATGTGAGCTTGAGAAGAATATATATTCTGCTGTTATTGGATAAAGTAGTCTACAGATATCAATCATATC
CAGCTGATTGACGGTGTTGTTGAGTTCAACTATGTCCTTACTGATTTTCCAGTTGCTGAATTTGTCCATTTCTGATAGAG
GGGTGCTGAAGTCTTCAACTATAACAGTAGATTCATTTATTTCTCCCAACAGTTCTTTCAGTTTTTGCCTCCTATATTTT
ATTGATCTGTTGTTAGGCAAATATACATTAAAGATTGTTATGTCTCTTGCAGAACTGACTCCTTTATCATTACGTAATAC
CCTTCTTTATCCTTGATAACTTTCCTTGCTTTGCATTCTGCTCTATCAGAAATTAATACAGCTGGCCGGGTGCGGTGGCT
CACGCCTGTAATCCCAACACTTTGGGAGGCCAAGGCAGGCAGATCACTTGAGGTCAGGAGTTCGAGACCGGCCTGGCCAA
CATGGCAAAACCTTATCTCTACTAAAAATAAAAAAAACTAGCTAGGCATGGCAGCACGTGCCTGTAATCCCAGCTACTCA
GGAGGCTGAGGCAGGAGAATCGCTTGAGGTTGCAGTGAGCTGAGATTGCACCACTGCACTCCAGTCTGGATGACAGAGCG
AGACTCTGTCAGAAAGAGAGAAAGAGAGAGAGAGAAAGAGAGATAAAGAGAGAAAGAGAGAGAGAAAGAAGAGAGA
GCAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAAAGAAAGA
GAAAGAAAGAAAGAAAGAAAAAAGAAGCAAGCTAATTCTGCTTTGTTTTGCTTAGGACTGGCATCCATCCAATTACTTTTA
ATCTATATGTGTCTTTATATTTAAAGTGGATTTCTTGGCTGGGTGCGGTGGCTCATGCCTGTAATCCCAGCACTTGGGAG
GCCGAGGTGGATGGATCGTCCGAGGTCAGGAGTTCGAGATCAGCCTGGCCAACATGGTGAAACCCCGTCTCCAATAAAAG
TACAAAAATTAGCTGGGCGCGGTGGTGGGCGCCTGTAATCCCATCTACTTGGGAGGCTGAGGCAGGAGAATTGTTGAACT
CGGAAGGCAGAGGTTGCAGTGAGCCGAGATTGCACCACTGCACTCCAGCCTGAGCCACAGAATGAGACTCTGTCTCAAGA
AAAAGTGGACTTCTTGTAGACAACATATAGTTTCATTGGGTCTTATTTTTCTTATTCACTCTAACAATCTCTGTCTTCCA
ATTGGTGCATTTGGAGCATTGATATTCACAATATTTATTAATACAGTTGGATTAATATCTACCGAATTTGTTGCTCTTGT
TCTTTGTTCCTATTTTTGTTT
```

Order-0 Markov model of the above sequence. An order-0 Markov model is the probability of occurrence of each nucleotide in the sequence. There is no positional correlation information in the 0th order Markov model.

A	C	G	T
0.224654926226	0.29128986197	0.303188957639	0.180866254165

Order-1 Markov model for the same sequence. Here, each row represents the base at any given position in the sequence. Each column of this row represents probabilities for the base in the *following* position. For example, for this sequence, the probability that an `A` will be followed by another `A` is 0.3366, and the probability that a `C` will be followed by a `G` is 0.06.

	A	C	G	T
A	0.33660130719	0.117647058824	0.303921568627	0.241830065359
C	0.318421052632	0.234210526316	0.0605263157895	0.386842105263
G	0.341101694915	0.209745762712	0.209745762712	0.239406779661
T	0.193396226415	0.188679245283	0.25786163522	0.360062893082

Order-2 Markov model for the same sequence. Rows correspond to all dinucleotide possibilities, and columns correspond to all four possibilities for the following base. Thus, given the two preceding bases (i.e., row), the columns represent probabilities for the base in the *following* position. For example, the probability of an `G` to follow `TC` is 0.05.

	A	C	G	T
AA	0.45145631068	0.0922330097087	0.26213592233	0.194174757282
AC	0.333333333333	0.194444444444	0.0555555555556	0.416666666667
AG	0.47311827957	0.155913978495	0.198924731183	0.172043010753
AT	0.216216216216	0.189189189189	0.195945945946	0.398648648649
CA	0.214876033058	0.173553719008	0.396694214876	0.214876033058
CC	0.415730337079	0.202247191011	0.0786516853933	0.303370786517
CG	0.347826086957	0.173913043478	0.304347826087	0.173913043478
CT	0.183673469388	0.210884353741	0.272108843537	0.333333333333
GA	0.316770186335	0.0869565217391	0.39751552795	0.198757763975
GC	0.30303030303	0.262626262626	0.0606060606061	0.373737373737
GG	0.282828282828	0.30303030303	0.131313131313	0.282828282828
GT	0.16814159292	0.176991150442	0.29203539823	0.362831858407
TA	0.292682926829	0.146341463415	0.162601626016	0.39837398374
TC	0.25	0.258333333333	0.05	0.441666666667
TG	0.225609756098	0.219512195122	0.256097560976	0.298780487805
TT	0.197368421053	0.179824561404	0.271929824561	0.350877192982

Figure 3.1: Markov models of a randomly picked sequence.

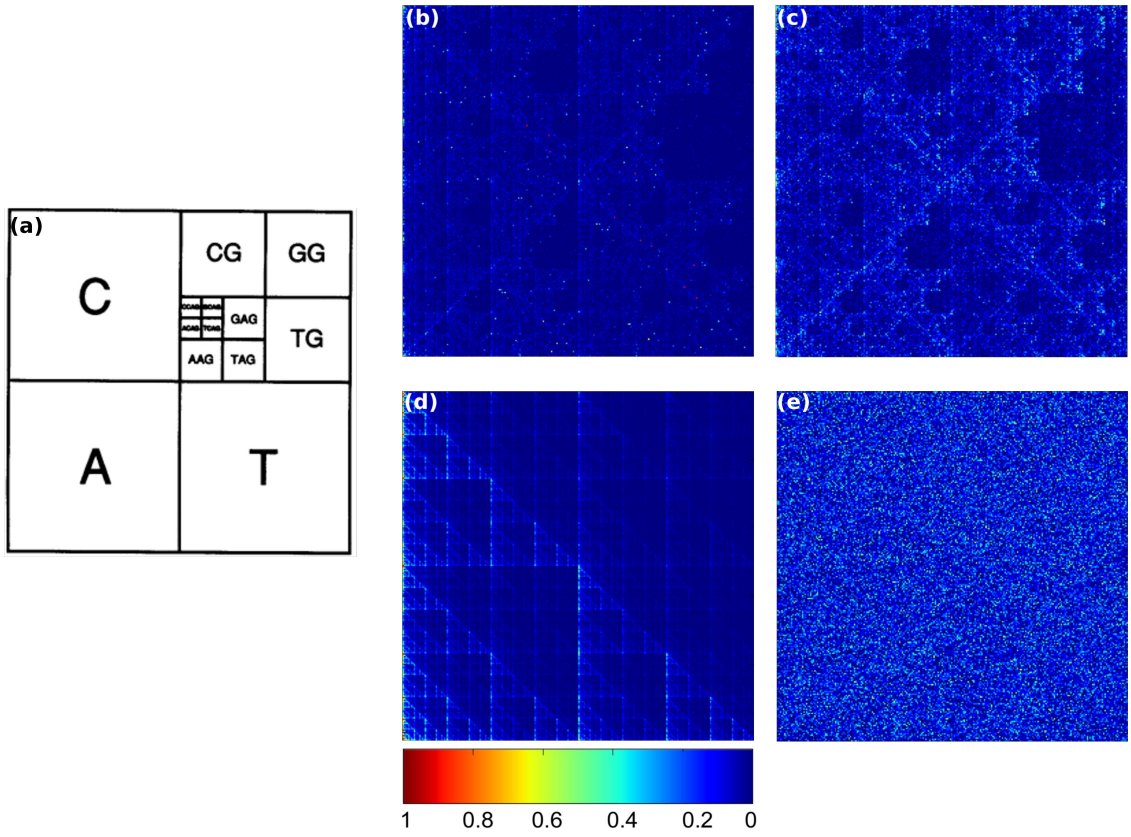


Figure 3.2: **Construction of the CGR**: A CGR organizes oligomers of fixed length in a hierarchical fashion over a square. The CGR construction is illustrative in panel (a) (figure adopted from Goldman (20)). A fCGR is a color-coded representation of the relative *frequency*-CGR (i.e., frequencies of oligomers normalized by the highest frequency). Color bar at the bottom is the key to these relative frequency values. Panels (b)–(e) are 8-mer fCGRs for sequence data from the following sources: (b) sequences picked from random locations in the human genome. (c) sequences generated by `RSAT` (21) using the 8th order Markov model. (d) sequences generated randomly assuming 60% GC content. (e) sequences generated randomly assuming equal probabilities for all the bases.

An initial application of CGR demonstrated that genes and intergenic regions can be differentiated visually from their CGRs (22). It was later demonstrated that the CGR alone is able to classify prokaryotes down to the genus using randomly-picked fragments of their genomes (or whole genomes when small) (23, 24). Furthermore, the information from CGR has also been shown to be useful in whole genome analysis (25). This is interesting because conventional sequence comparison approaches (such as Multiple Sequence Alignment (26)) assume that sequences being compared come from a specific region (e.g., 16s rDNA). This is yet another illustration of the usefulness of Markov models considering the one-to-one correspondence between a Markov model and the CGR (27).

As a concrete example of the utility of the CGR, we have plotted (Figure 3.2) the frequency CGR (fCGR) for DNA sequence data from a variety of sources. Frequencies plotted are *relative* frequencies normalized by the maximum frequency. In other words, to plot a fCGR, the frequencies of oligomers are normalized by the frequency of the most occurring oligomer, and color coded accordingly. The color bar in the figure is the key to the relative frequency value. Panel (a) shows a schematic used to construct CGR as illustrated by Goldman (20). Each fCGR is plotted at a resolution of 8 mers. In Panel (b), the fCGR is plotted for sequences downloaded from random locations in the human genome. In Panel (c), the fCGR is plotted for sequences generated using a 7th order Markov model in `RSAT` (21). This model was generated from a single contig in the human genome (21) (personal communication). In Panel (d), the fCGR is plotted for sequences generated assuming a 60% `GC` content. In Panel (e), the fCGR is plotted for sequences assuming equiprobability for all the nucleotides. Thus in the Figure 3.2 panels (b) through (e) are made of exactly 4^8 i.e. 65,536 points plotted using the scheme as illustrated in panel (e).

It can be seen that in the CGR of sequences picked randomly from human genome (panel b) and CGR for a sequence generated from order-7 Markov model (panel c) appear similar. However, on careful examination it can be seen that the CGR of sequences from the human genome have a more well-defined structure as compared to the CGR of sequences generated from the order-7 Markov model. When sequences are generated randomly, assuming independence of successive nucleotides and a probability distribution ($p_A = p_T = 0.2, p_G = p_C = 0.3$) corresponding to a total GC content of 60%, the structure of the corresponding CGR (panel

d) is not as complex as that in panel b. The CGR of IID nucleotides with the probability distribution ($p_A = p_T = p_G = p_C = 0.25$), as expected, has no discernable structure.

These examples illustrate that fCGR is a useful tool for visualizing the structure of and the nature of sequential correlations in DNA sequences.

3.1.2 Survey of Available Motif Search Algorithms

In early 90s there were vast improvements in the DNA sequencing technology making it available to a number of establishments in the world. When the amount of sequence data was very small the comparisons could be done manually. As the amount of sequence data available in the public domain increased exponentially it became infeasible to analyze these sequences manually. This evolution in sequencing technology coupled with simultaneous improvements in computers made increasingly advanced computers a perfect tool for analyzing the vast sequence data being produced. Most of the tools described here were primarily created to analyze protein sequences. However over the number of years most these algorithms and programs have been adapted to work with DNA and sequences as well. Hudak and McClure have compared various algorithms for motif detection using protein sequences as test data (28).

One of the earliest motif detection algorithms to be published was the `BLOCKMAKER` (29). Briefly, the algorithm searches for words consisting of codons (trimer) in a dataset and adjacent codons with highest frequencies are combined motifs (30).

Another algorithm uses the iterative protocol as described by Karlin & Broccheiri and is known as `ITERALIGN` (31). The method produces alignment blocks that accommodate indels and are separated by variable-length unaligned segments (31). The blocks are derived from the alignment of the consensus sequences and are improved by displacement of individual sequences. The blocks are defined by a consensus residue and conservation index.

`PIMA` (Pattern-induced multi-sequence alignment) is an algorithm that uses secondary structure dependent gap penalties in multiple sequence alignment (32). As this algorithm uses more information than the sequence itself it is able to to accurately align structural boundaries in a set of homologous sequences.

PROBE is another method that involves multiple sequence alignment and model generation using single short sequences (33). PROBE constructs an alignment model of the protein family through a combination of Gibbs sampling, a genetic algorithm database searches using progressively more refined alignment models (33).

The MEME is another well-known, well-cited and well-respected algorithm for motif finding in DNA and protein sequences (34). This algorithm uses a mixture model for determining the motifs and their positions in the training sequences (12), and uses the *Expectation-Maximization* (EM) method for maximum likelihood estimation in the context of incomplete (missing) data problems (10).

SAM is another algorithm based on Gibbs sampling (35, 36).

Approximate matches using 'Edit distance' have been proposed by Wang *et al.* (37). In this algorithm a consensus sequence is arrived at from related protein sequences. This algorithm works in conjunction with the BLOCKS algorithm described by Henikoff & Henikoff (38).

3.1.3 Confounds in Motif Detection

In addition to the complexity and the long-range correlations in the genomic sequences, any motif detection exercise needs to worry about confounds – specifically, the repeat sequences present in the genome.

It is generally believed that only 2% of of the entire human genome consists of coding regions, i.e., genes that code for a protein/RNA. The rest of the genome was considered to be made of "junk" DNA and the understanding of the function of this majority of genomic DNA is only emerging now (39, 40). These non-coding/intergenic or ill-understood regions are known to be made up, predominantly, of repetitive stretches of DNA. For a note on repeat families and especially the *Alu*-repeats see Chapter 2. It is believed that repeats play important role in many biological processes (e.g., retroviral integration (3), heterochromatinization (41), etc.) and thus cannot be ignored completely from study. Similarly a large portion of the genomic DNA is also made up of *satellite* DNA (42, 43), which has a very specific function and plays important role in maintaining the structural integrity of the genome during replication. These sequences may or may not be part of relevant motifs when such an

exercise is undertaken to study role of motifs in regulation of gene expression. The function of many repeat sequences still remains unknown.

Any such repetitive element that occurs with high propensity in the data being analyzed tends to be picked up as a motif because of the very nature of probabilistic motif detection algorithms. This is an undesirable outcome of a motif detection analyses if the purpose not the discovery of such repetitive elements itself. It is a common practice to minimize the effect of such repetitive elements by *masking* the repetitive element(s) during analysis such that these part(s) of the sequence(s) are replaced by a single character. When DNA sequences are masked the replacing character is usually `N` and it is `X` when protein sequences are masked.

Fundamentally, all present day probabilistic motif detection algorithms lack the ability to distinguish motifs from repeats because that would necessitate a fundamental revision of probabilistic models of genomic sequences to include statistically distinguishable models of repeats, and appropriate modifications in the motif detection formalisms. We believe that this may be a difficult task, and may lead to a steep increase in the computational demand of such a methodology.

3.1.4 The Background Model: Why Is It So Important?

At the heart of all probabilistic motif search methodologies lies such a probabilistic model of biological sequences. This model usually comprises of a *background* part and a *motif* part. A motif is usually modeled as a sequence of independent nucleotides with probability distributions that depend on the position of the nucleotide within motif (see Figure 3.1 an example). On the other hand, the background is usually modeled using a Markov model of appropriate order.

It is a common practice (in single-species analyses) to choose the order and the parameters of the background model so as to match some desirable characteristic of the genome (as represented, usually, through a set of background sequences) to which the sequence data being analyzed belongs. For example, requiring a match on the average genome-wide %`GC` content alone corresponds to an order-0 model, while matching the average genome-wide k-mer frequency profile corresponds to an order-$(k-1)$ Markov model. In the worst-case scenario, parameters of the background model can be estimated using the same sequence

data that is being analyzed for motifs as in the default background setting for `MEME` (12). While the order-0 Markov model (i.e., independent nucleotides identically distributed (IID) according to pre-specified probabilities) appears to have been employed extensively in the early days of motif search (see, e.g., (44)), recent reports recommend the use of an order-3 Markov model for the same purpose (13, 45).

It could be argued that the general principle governing the choice of the background should be as follows: In the context of the scientific problem being investigated, the background should be statistically similar in all respects to the sequence data being analyzed except, possibly, for the feature of interest (i.e., the motif) (46). More precisely, while the choice of sequence data for motif search is governed by the scientific question being investigated (i.e., the alternate hypothesis), the appropriateness and choice of the background is dictated by the corresponding null hypothesis (that the sequence data and the background are statistically indistinguishable).

Choosing parameters of the background model (alternatively, an appropriate set of background sequences from which the background model could be built) correctly is of vital importance to any motif search exercise. An improper choice of the background can lead to unreasonable numbers of false positive/negative results or biologically irrelevant motifs. This greatly diminishes the value and reliability of the conclusions drawn from such analyses: Indeed, motif discovery methods have been criticized (47, 48) on the ground that they tend to report false positives.

A comprehensive review of literature on the problem of background selection appears in Marchal *et al.* (49) with a focus on motif detection in prokaryotes. Thijs *et al.* (13) reports extensive in silico experiments, in the context of the *Arabidopsis* genome, on the effect of the background model on the quality of motifs detected. The key observations and recommendations of these two works taken together are as follows:

1. It is essential to use background models generated from genomic DNA so as to increase the performance of motif detection algorithm.

   Here, the term *performance* implies biological relevance and statistical significance of the detected motifs.

2. It is desirable to use a background model of as high an order as possible.

Practically, however, it becomes virtually impossible to generate models with order $n \geq 6$. This is because an enormous amount of background sequence data is usually required to get adequate representation of all the 4^{n+1} n-mer frequencies used to estimate model parameters.

Specifically, pseudo-counts added to compensate for unobserved occurrences of oligomers not represented in the sequence data usually leads to deterioration of the quality of detected motifs.

3. Order-3 or higher Markov models generated from the input sequence data for motif detection itself, when used as background, severely hamper the ability of motif finding algorithms to detect valid motif(s).

Apart from these two extensive studies, there exist no useful guidelines in the literature for the choice of the background model order (especially for a complex genome such as the human genome), to the best of our knowledge.

3.1.5 Motivation for the Present Work

We wish to investigate the effect of choice of background model on the biological relevance and usefulness of the discovered motifs in the setting of the human genome. As a concrete biological scenario for studying the role of the background on the quality of motifs detected, we have chosen the problem of detecting motifs in and around HIV integration sites. This problem has been extensively discussed in Chapter 2. Indeed, the work presented here has its origins in the work presented in Chapter 2.

We first provide a brief summary of this problem of motif detection in HIV integration target sites in the next section, in the context of the present work. In the following section, we formulate our choice of the proper background for this problem.

3.2 Motifs in HIV Integration Sites

The problem of how the HIV selects integration sites in the human genome is an important open problem in virology and human pathology. An understanding of this process will not only shed light on the retroviral integration site selection process, it will also help design

effective retroviral vectors for gene therapy. It is known that integration site selection not only affects the ability of the virus to complete its own lifecycle (41), it also affects the host (50). Let us briefly review what is known about HIV integration site selection; a detailed discussion can be found in (51, and references therein) and Chapter 2.

HIV is a RNA virus that infects human T-cells and causes AIDS (Acquired Immuno-Deficiency Syndrome) follows an intricate and highly regulated lifecycle. Briefly, HIV injects its genomic RNA into the host cell. The genomic RNA of the virus is reverse transcribed into cDNA, and is also translated to produce a few virus-specific proteins. This cDNA, virus-encoded proteins, and some host-encoded proteins together form pre-integration complexes (PICs) (52, and references therein). The PICs are then translocated to the nucleus where they bring about integration of the viral cDNA into the host genome.

It was believed (53), until Schroeder *et al.* published their paper (54), that HIV integration sites are randomly dispersed within the human genome. However, it was also known that genomic locations where the viral cDNA integrates into the host genome control the fate of the virus (for example, it was known that specific highly heterochromatic regions in the genome (such as centromeric alphoid repeats (41)) are avidly avoided by the PICs). The role of integrase in integration site selection had also been underlined (55).

Schroeder *et al.* (54) demonstrated that the distribution of HIV integration sites in the human genome is far from random, and moreover, it is positively correlated with the gene density on a chromosome. Later, it was also shown that different retroviruses have distinct preferences for integration sites in their respective target genomes (56, 57). Using `GenBank`-deposited sequences from previous studies (54) it was shown (58) that there are definite base preferences for retroviral integration. In particular, HIV-1 shows a preference for symmetric bases in the target genome (59). Our own previous work demonstrated that HIV integration target sites are characterized by their unique chromatin context as defined by specific consensus motifs. In passing, we note here that despite all these analyses, the nature of HIV integration target sites needs further investigation for better characterization (3).

3.2.1 The Ideal Background for Motif Search in HIV Integration Data

Clearly, the primary bioinformatic tool for an *in silico* analysis of this problem of analyzing the integration sites of the HIV in the human genome is motif detection. Indeed, all the works mentioned in the previous paragraph involved a substantial bioinformatics component. In the light of the discussion in Sec. 3.1.5, it may be said that the null hypothesis in this context would be *"HIV integration sites are random locations in the human genome"*, whereas the alternate hypothesis would be *"HIV targets specific, non-random locations in the human genome for integration"*. Motif detection, in this context, is an attempt to weigh, albeit indirectly, these two hypotheses against one another for a given dataset. In addition, it attempts to characterize the integration sites by searching for consistent patterns specific to the integration regions.

This formulation of the basic scientific problem under investigation clearly suggests that the most appropriate background for the present motif search exercise would be a set of sequences (of suitable length) picked from *random locations distributed uniformly over the entire human genome*.

3.3 Materials and Methods

3.3.1 Overview of Analysis Protocol

This study was performed in three parts. Below we provide a quick overview of these three parts, and the rest of this section elaborates upon their details.

1. Data preparation • This involved retrieval and pre-processing the HIV integration site sequence data, retrieval and pre-processing of the background sequence data, building Markov models (of orders upto 6) of the background sequence data, and sensitivity analysis of the estimated model parameters.

2. Motif detection • We performed an extensive motif detection exercise for the HIV integration site sequence data against background models of orders upto 6 thus constructed. For each background model, we used `MEME` (60) to detect 5 non-overlapping best motifs with allowed length variation between 5 and 50.

Background	Description
MD-0	Order-0 Markov model generated from the input sequence data itself. This is the `MEME` default background.
RP-0	Order-0Markov model built from 1000 sequences of length 10000 bases each, picked from random locations in the human genome.
RP-1	Order-1 Markov model built using the above set of randomly-picked sequences.
RP-2	Order-2 Markov model built using the above set of randomly-picked sequences.
RP-3	Order-3 Markov model built using the above set of randomly-picked sequences.
RP-4	Order-4 Markov model built using the above set of randomly-picked sequences.
RP-5	Order-5 Markov model built using the above set of randomly-picked sequences.

Table 3.2: Description of background models used for our motif detection excercises.

We also performed a similar exercise using the `MEME` default background model which, in our case, was an order-0 Markov model constructed from the HIV integration site sequence data itself. In Table 3.2 we summarize the various background models used and what they mean.

3. Assessment of biological relevance • For each motif thus detected, we located its occurrences in the `Genomatix` promoter database using a profile match, and obtained the corresponding gene information. For each order of the background model, we performed a hierarchical `GO`-Term enrichment analysis of the corresponding genes thus obtained, using `AmiGO` (61).

   This analysis leads to the key result of this work that in a motif exercise, a properly chosen background model leads to biologically more relevant and meaningful results.

3.3.2 Data Preparation

HIV Integration Sequence Data

For the second part we used published sequences for HIV integration sites (54) and obtained their flanking genomic sequences. The process of obtaining genomic sequences for our analysis is described in detail in Chapter 2. We give an overview of the process here for the sake of completeness.

Briefly, each submitted sequence from Schroeder *et al.* was compared with the NCBI `RefSeq` under default settings[1]. Of the matching regions from the genome the best match was selected (the match with maximum score and minimum E-value). If there were multiple matches with identical scores and evalues the first match was chosen. From each of the matches we generated genomic coordinates such that the match is in the center of a 2 kb genomic region. These sequences were downloaded and became our starting point.

These sequences were processed through a locally installed copy of the `CENSOR` (62). From the masked sequences the extent of repeats present in each sequence were calculated. These sequences were 'weighted' inversely to the extent of repeats present such that the most masked sequence(s) got least weight and least masked sequences got maximum weight. Thus a sequence that contained no repetitive elements got a weightage of 1.0 and the sequence that was completely masked got a weightage of 0.005 (`MEME` cannot accept sequences with weight 0). These sequences became our input sequences for the motif detection exercise using the `MEME` (63).

Background Data, Models, Stability

Background Sequence Data. As discussed earlier, the most appropriate background in the context of the present work consists of sequences picked from the human genome from random locations. Such randomly-picked sequences can then be used to build Markov models of any reasonable order. Using the `Random Sequence Grabber` tool (described below), we downloaded 1000 sequences of length 10000 bases each, from random locations in the human genome (build 36 v3 of the `RefSeq` at the NCBI). These sequences were then used to generate Markov models of order 0 through 6 using `GenRGenS` (64).

[1] `http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid=9606`

The `Random Sequence Grabber` Tool. To the best of our knowledge, no tool/software/ program that enables downloading of sequences from random locations in a given genome is available in the public domain. We thus designed and created a tool, called the `Random Sequence Grabber`[2], for this purpose. The structure of this tool is illustrated in Figure 3.3. Briefly, a given genome is considered to be a single linear molecule starting with first nucleotide of the first chromosome, and ending with the last nucleotide of the last chromosome. This number is very large, and there are not many random number generators that can handle such a large range. The computing environment/programming/scripting language called `python`[3], however, comes with an in-built implementation of the **Mersenne Twister** random number generator (65) that has a very large period of 2^{19937}, and is thus perfectly suited the purposes of this tool.

In the `RefSeq` database of the NCBI, the reference sequences for many genomes are stored in chromosome-wise fashion. Each chromosome is in turn made of number of contigs. There are almost always some non-sequenced regions in the chromosomes which show up as 'Gap'. Thus, a chromosome can be represented as a set of contigs separated by gaps. The length of each contig and each gap is known. Using this information, a cumulative length table can be constructed, such that each number corresponds uniquely to a *contig* or a *gap*.

The program generates n (predefined) random numbers. Each random number is checked against cumulative lengths of contigs. The nearest bigger cumulative length is chosen from the hash table. The program then looks for the information defined by that length. If the number defines a *Gap*, a new number is chosen. If the number defines a *contig*, then start s, and end e coordinates of the sequence to be downloaded are determined. This depends on the length l of the sequence to be downloaded. It is ensured that the final coordinates of the sequence lie within the same contig. If this condition is not met, a new number is chosen. Coordinates are generated by adding and subtracting (length / 2) from the number. These coordinates are used to generate the download URL. These URLs are written a to a temporary file, and if a internet connection is available, then these sequences are downloaded one at a time.

[2]Available at the `http://sourceforge.net/projects/genome-seq-grab/`. The source code is also available under the GNU-GPL version 3.0 or later

[3]`http://www.python.org`

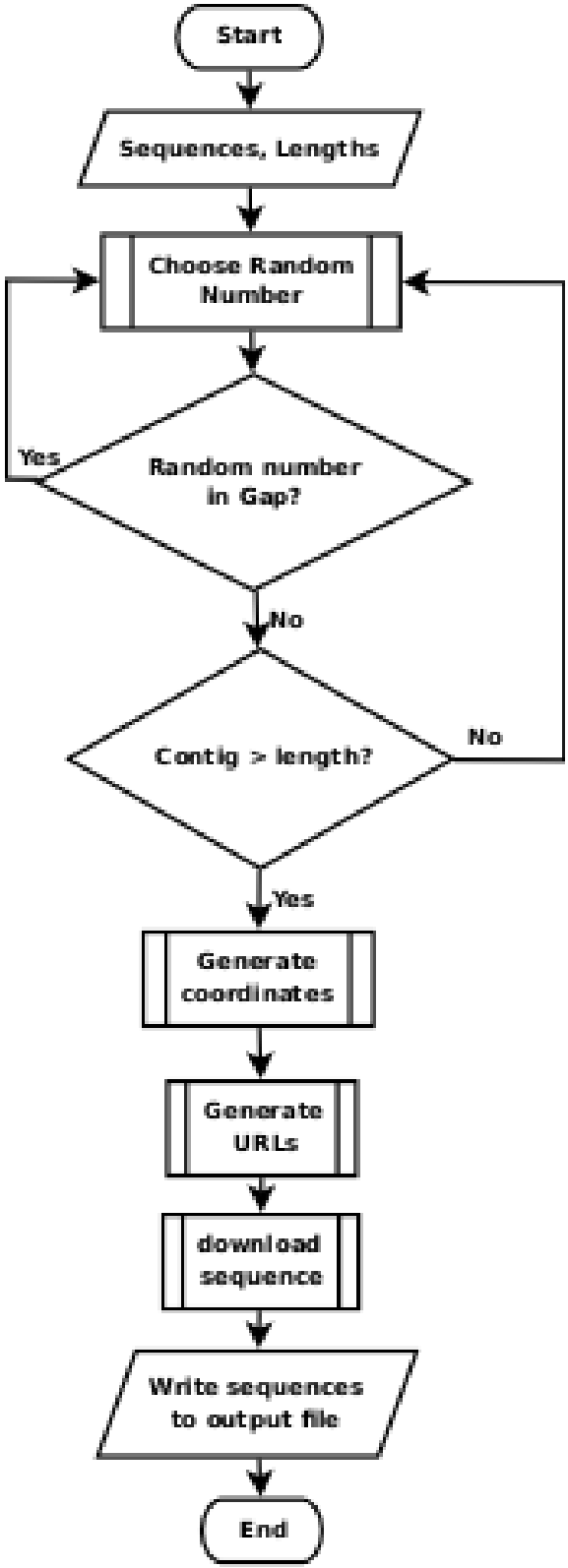


Figure 3.3: Algorithm used to pick a sequence randomly from the genome.

A download URL is created using the generated coordinates and the template url[4] such that the sequences are downloaded in the `FASTA` format. It is in principle possible to download the sequences in `GenBank`, `XML`, `ASN.1` or any other standard format as available at the NCBI. We chose the `FASTA` format because it is the most commonly used input sequence in many motif detection algorithms.

Following Figure 3.4 shows the screen shot of the Random Sequence Grabber GUI. The interface is very simple on purpose, because it is meant to be used by biologists who should not be bothered with knowing the intricacies of programming.

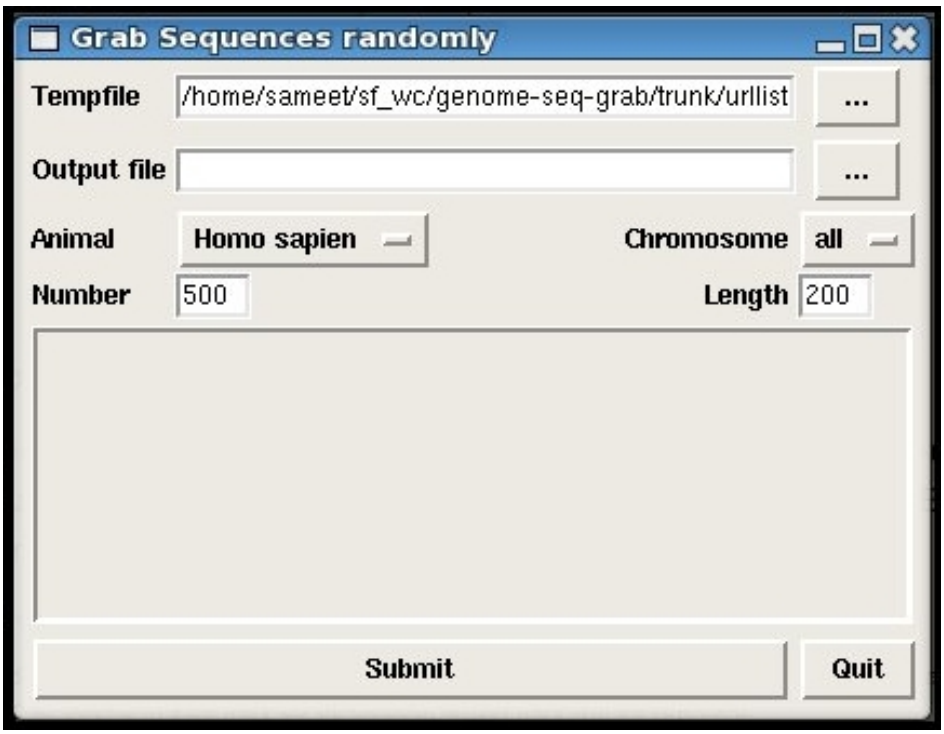


Figure 3.4: Screen shot of the random sequence grabber program.

Validation of the `Random Sequence Grabber` tool. We used uniform distribution to generate the random number. It therefore stands to logic that under uniformly generated random numbers the proportion of sequences being picked will correspond to the lengths of the chromosomes. That is precisely what is seen here. Table 3.3 shows the correlation between the proportion of sequences drawn from a chromosome and length of the chromosome with re-

[4]There are many ways to generate such a template url, most of these are available at `http://www.ncbi.nlm.nih.gov/projects/mapview`. Moreover, such urls can be inferred from all places at NCBI that allow "download sequence" feature.

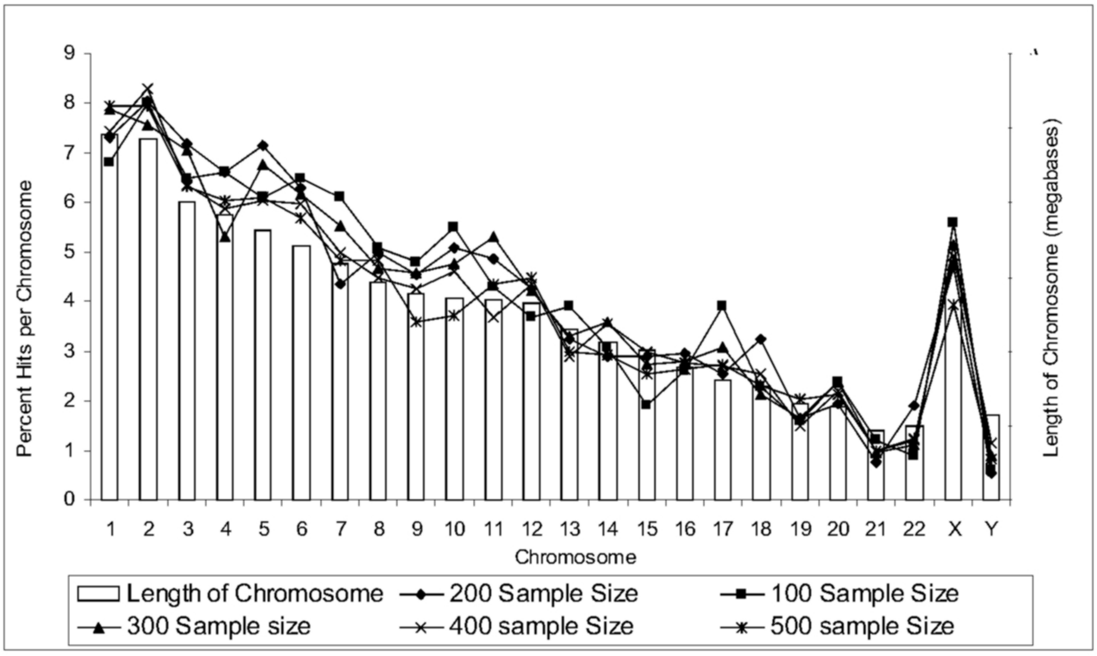


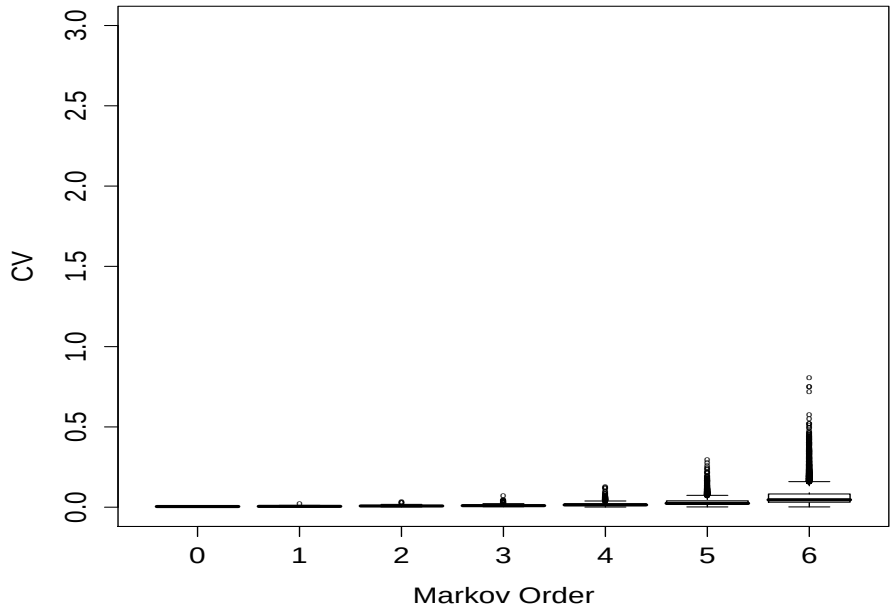
Figure 3.5: Number of sequences picked randomly show a distribution that reflects length of the chromosomes in the human genome.

spect to the sample size n when the sequences are drawn from the human genome. As the size of sample n drawn increases the proportion of the sequences picked closely resembles the length of the sequences.

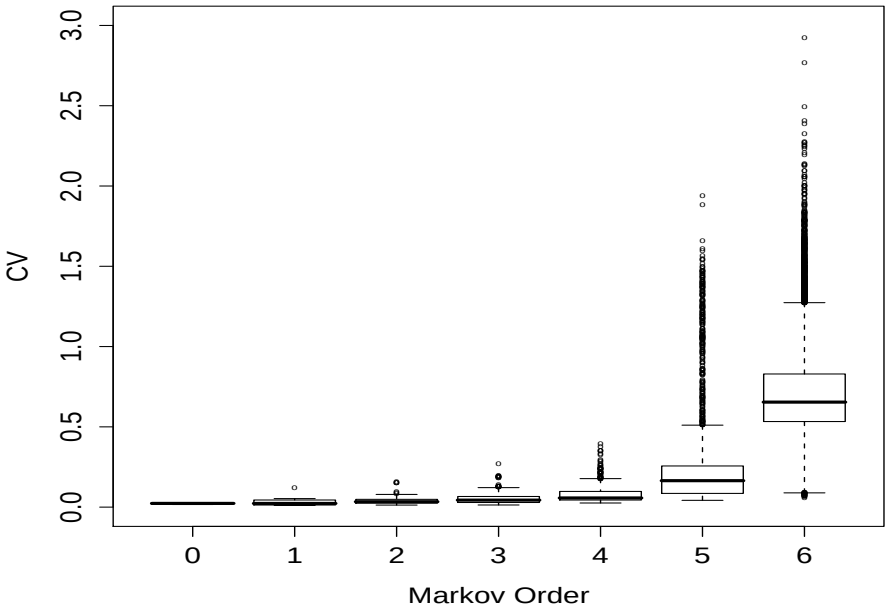
This test establishes that the program to download sequences randomly from the human genome is downloading sequences proportional to the length of the chromosomes, and that the sequences picked span the entire genome.

Robustness of the Background Models. As mentioned before, we downloaded 1000 sequences, each of length 10000 bases, from the human genome (NCBI `RefSeq` build 36 v3) using the `Random Sequence Grabber` tool mentioned in the previous section. These sequences were used to build Markov models (using the `GenRGenS` (64) tool) of various orders for use as background for the motif detection exercise.

It is desirable that the results of a motif detection exercise do not depend sensitively on the specific set of background sequences used to build background models. Such sensitivity of estimated Markov model parameters can be assessed by constructing multiple models instances for each order using entirely different sets of randomly-picked sequences. To this



(a)



(b)

Figure 3.6: **Variability of Markov model parameters over multiple background sequence sets**: What is shown is the distribution of the coefficient of variation (CV, Equation 3.1) in the form of a boxplot, as a function of the model order: (a) CV variability for Markov models built using 5 independent sets consisting of 1000 sequences length 10000 each, and (b) CV variability for Markov models built using 50 independent sets consisting of 100 sequences length 10000 each.

The Boxplot is a representation of data in a way such that the overall distribution of the values in the data can be observed in a single figure. The two *staples*, one at the top and the bottom denote the maximum and minimum data values. The lower (upper) border of the box denotes the 25th (75th) quantile of the data, the thick line in the middle of the box denotes the median of the data, and data points designated as outliers are shown as small circles.

Sample size (n)	**Correlation coefficient (r)**
100	0.7907
200	0.8019
300	0.8771
400	0.8948
500	0.9568

Table 3.3: As the sample size increases the proportion of the sequences picked from different chromosomes starts approaching proportion of the length contributed by each chromosome under the assumption that that a genome is a single molecule with all the chromosomes joined end to end.

end, we downloaded 5 independent sets, each with 1000 sequences of length 10000 each. Each set was used to build Markov models of orders 0 through 6. We then assessed the variability of the Markov model parameters via the distribution of their coefficients of variation (CV), defined as the ratio of estimated standard error σ of each model parameter to the estimated mean μ of each model parameter:

$$\text{CV} = \frac{\sigma}{\mu} \tag{3.1}$$

Figure 3.6 shows boxplots of the CV as a function of model order. These boxplots show very clearly that the CV values for a very large number of models parameters (an order-k model is defined by 4^{k+1} parameters) are rather small. This implies a small variability of only a few percent of the mean value. We also see a mild increase in the variability of parameter values with the order of the model, as seen through the variation of the median CV value as a function of order. Each of the outliers seen in the plot corresponds the CV values for *one* of the model parameters; the number of such outliers is clearly quite small (Figure 3.6(a)). This implies that Markov models built using randomly-picked sequences of sufficient length (10000 bases) and number (1000) are quite stable, and not very sensitive to the specific instance of sequence data used to build them.

To assess the stability of model parameters built from a smaller-sized sequence set, we also downloaded 50 independent sets, each with 100 sequences of length 10000. Again, each of these 50 sets was used to build Markov models of orders 0 through 6. In Figure 3.6(b) we again depict CV variation for this set of models, as a function of the model order. We clearly see that order-5 and order-6 Markov model parameters show wide variability. This is in contradiction with the popular perception that 100 sequences of length 10000 each is a

fairly large set of genomic sequences. This variability of model parameters is traced to the observation that not all 6- and 7-mers were represented in many of these smaller sets, thus making it difficult to estimate stable Markov model parameters for order 5 and 6 respectively (See page 87 for detailed explanation). The important insight gained through this exercise is twofolds:

1. Background model variability needs to be assessed to ensure stability of motifs detected, and
2. Building stable Markov models of high orders needs sufficiently large sequence sets. Clearly, this need will grow exponentially with respect to the intended Markov model order.

3.3.3 Motif Detection

We used the `MEME` (63) tool extensively for our motif detection exercises. `MEME` is a very highly computation-intensive algorithm. For the present work, the high-performance server used had two 64-bit dual-core Intel `Xeon` processors and 12 GB of RAM, with `Linux (Fedora 8)` as the OS. A local copy of `MEME` (version 3.5.4) was compiled in serial mode using Intel compilers (version 10.1). For the HIV integration site data described earlier (See page 99), `MEME` run took about 220 hours (about 9 days) to find 5 motifs with each of the 6 background models (i.e., orders 0 through 5 Markov models constructed from sequences randomly picked from the human genome, as described earlier (See page 87).

The command line for a typical `MEME` run had the following structure. Specific command-line arguments used in this command line are explained below. Further details on the usage of `MEME` can be found in (63, 66) (also see the `MEME` website `http://meme.sdsc.edu` for more details).

```
meme invivo_sequences.txt -dna -mod tcm -nmotifs 5 -nsites 372 -wnsites 0.8 -text -bfile human_random_order3 \
-maxsize 1000000 -revcomp -text -minw 5 -maxw 50 > invivo_motifs_order3_background.txt
```

Model This is the `-mod` option in the `MEME`. It can take 3 values viz., ZOOPS, OOPS, and tcm. We used the option `tcm` because we expected zero or one or more repetitions of the motifs per sequences. The value of ZOOPS only looks for zero or one occurrence of

motif per sequence, and the value of OOPS looks for exactly one occurrence of motif per sequence.

Number of motifs This is the `nmotifs` option. The value is number of motifs that should be searched. We searched for 5 motifs over each background.

Background This option is given as `-bfile`. The value to this option is expected to be a file with specific structure (n-mer probabilities).

We constructed the background files from randomly picked sequences as mentioned earlier and used backgrounds of order from 0 through 6 in addition to the default background of `MEME` while all other parameters remained the same.

Number of sites `-nsites`, the value of this option determines number of times a given motif should occur to qualify as a motif under given parameters. The default value for this option is 50. We used `nsites`=Number of sequences.

Data size `-maxsize`, this argument is an estimate of total data that will be processed by the `MEME`. The default value is 100000, however the data used in the current study is of size 372 (number of sequences) $\times$2000 (length of each sequence) so we changed `-maxsize` to 1000000.

Strands `-revcomp`, arguments considers the input sequence(s) and their reverse complements separately.

3.3.4 Assessment of Biological Relevance

It is known that retroviruses prefer integration in transcriptionally active regions of chromatin such as promoters and first introns (67). In the context of the present problem, viz., motifs in HIV integration sequences, it thus makes imminent sense to look for the occurrences of detected motifs in *promoter* sequences. Biological relevance of detected motifs can then be assessed by the strength of their association with the HIV life cycle.

Our protocol for the assessment of biological relevance of detected motifs as function of the background model order is summarized below; detailed discussion of each step follows.

1. Using a profile match using a home-brewn tool (described below in detail), we located all occurrences of all the detected motifs in the human promoters in the `Genomatix`

Promoter Database (GPD; discussed in detail below). We thus obtain, for each order of the background model, a set of genes whose promoters contain one or more instances of the corresponding motifs.

2. We performed a GO-term enrichment analysis in an hierarchical and incremental fashion over these orderwise sets of genes thus obtained. The biological relevance of motifs corresponding to a Markov model order can then be assessed from the enriched GO-terms and their putative role in the HIV life cycle.

Why GPD?

The Genomatix Promoter Database (GPD) is a high-quality commercial promoter sequence database, available at http://www.genomatix.de. Unlike, e.g., the Eukaryotic Promoter Database (EPD), the GPD is a more complete and better-curated database with about 25000 promoters, together with relevant information such as gene IDs, gene symbols, alternative names, etc., which made the further analysis much easier and manageable.

Motif Profile Match Method

Given the motif model for a motif of length L, our profile match method is as follows:

1. We first calculate, under the motif model, the probability of the motif itself: this is defined as the maximum possible probability p_{max} under the motif model for any sequence of length L. Given the structure of the motif models (as explained in Section 3.1), p_{max} is simply a product of the sitewise maximum probabilities.

2. Next, we scan all sequences in the given sequence data (i.e., GPD). For each sequence S, we compute the probability $p(s)$ of every length-L subsequence s of the sequence, under the motif model. All such subsequences with $p(s)/p_{max} \geq 0.9$ qualify as occurrences of the given motif, and we extract gene information corresponding to sequence S.

We performed this exercise for all the detected motifs for each background model order, and obtained sets of genes labeled by the background model order.

GO-Term Enrichment Analysis

The GO Project. The GO (Gene Ontology) project (61) has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. The advent of high-throughput methodologies (microarray, ChIP-chip, etc.) as necessitated the development of ability explain co-expression patterns of number of genes (68). In microarray experiments the genes are grouped according to their expression patterns and then the associated GO-terms are analyzed for enrichment. There is copious amount of literature is available today on usage, importance and analysis of the GO terms (69, and references therein). A complete discussion of the structure of the GO project is beyond the scope of this thesis; we refer the interested reader the GO project website `http://gene-ontology.org/` for further details.

The GO database is most commonly used for detecting significant over-representation of GO-terms associated with a set of genes of interest, with respect to a global set of genes that contains the genes of interest. For example, in an expression microarray study involving multiple time points, a set of genes with similar expression profile over the period of experiment could be the set of genes of interest, as against the complete set of genes represented on the microarray. Functional annotations of the genes of interest are derived by analyzing which of the GO-terms associated with these genes got enriched with respect to the global set of genes.

Qualitatively, a specific GO term is considered enriched if the proportion of associated genes in the set of interest happens to be significantly larger than that in the global set. Quantitatively, this is achieved using statistical tests based on the hypergeometric distribution, and the significance of enrichment is assessed using the corresponding p-value (70) possibly with Bonferroni correction for multiple testing (71, 8). A variety of tools for such GO-term enrichment analysis are available in the public domain; e.g., AmiGO from the GO Consortium (72), the packages GOstats and HyperGOstats under the R statistical computing environment (73), etc.

GO-Term Enrichment Analysis of Detected Motifs. Our assessment of the biological relevance of the motifs discovered in the HIV integration sequence data is based on the

following biologically reasonable idea: We consider a discovered motif highly relevant if it occurs in the promoter regions of genes that are intimately associated with HIV-life-cycle-related processes in the host. For a motif with high biological relevance to the HIV integration problem, we expect to see an enrichment of GO-terms associated with such processes as immune response, cytokine secretion, etc. Specifically, we wish to assess how the order of the background model affects the biological relevance of detected motifs. This we do using GO-term enrichment analysis as follows:

For each motif pertaining to given background model order, we obtain list of associated genes using the profile match method explained earlier (page 109). Let G_k be the collective (i.e., across all motifs) list of all genes associated with background model orders $k = 0, 1, \ldots, n$, where n stands for the maximal order of the background model used for motif detection. Let

$$\begin{aligned} G &= \bigcup_{i=0}^{n} G_k \\ C &= \bigcap_{i=0}^{n} G_k \end{aligned} \tag{3.2}$$

be, respectively, the union and intersection of gene sets $G_k, k = 0, 1, \ldots, n$. The set G forms the global set for our GO-term enrichment analysis. The set C is the set of genes common across all orders, hence not relevant to analyzing the effect of the background model order on the biological relevance of motifs detected.

To extract the specific effect of the background model order, we we now form a ordered hierarchy of gene sets in two ways:

1. **Incremental Gene Set at Order k:** Define

$$\begin{aligned} g_0 &= G_0 - C \\ g_1 &= G_1 - g_0 \\ &\vdots \\ g_n &= G_n - g_{n-1}. \end{aligned} \tag{3.3}$$

Here, the operator '$-$' stands for the operation of set difference: For example, g_0 is the

set of all genes associated with order 0 that are not present in the set C of genes common across all orders, g_1 is the set of all genes associated with order 1 that are not present in the set g_0, and so on and so forth. Thus, the set g_0 is the set of genes that make their first appearance through motifs found at order 0. Clearly, some of these genes *may* show up at higher orders, but we consider these genes as being uniquely associated with order 0 because they show up, for the first time, with order 0 background model.

This construction is motivated by the observation that (a) the same motif may get detected at multiple background orders, and (b) there is no unique way of deciding whether two motifs, as characterized by their respective `PSPM`s, are identical. The above construction circumvents the problem of arriving at a unique set of motifs either within the set of motifs discovered over *one* specific background model order, or across multiple orders.

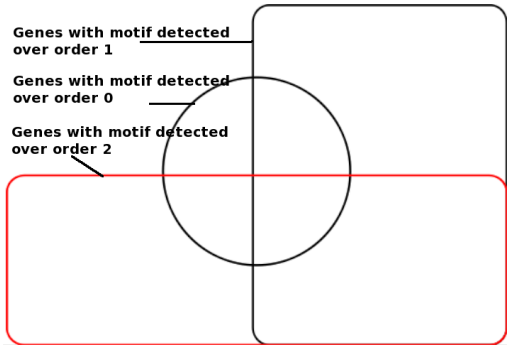
2. **Set of Unique Genes at Order k:** The set of genes uniquely associated with motifs discovered using order-k background model can be constructed as follows:

$$\gamma_k = G_k - \bigcup_{i=0 \ (i \neq k)}^{n} G_i \tag{3.4}$$

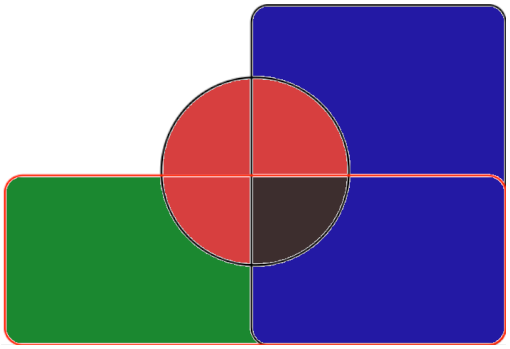
This construction is motivated by the need to assess effects specific to the given order.

These two constructions are illustrated schematically in Figure 3.7 for $n = 2$, using the Edwards representation of a Venn diagram (74). Figure 3.7(a) shows three sets, the black circle representing G_0, the black rectangle representing G_1, and the red rectangle representing G_2. Our global set G corresponds to the region enclosed by the three sets together. Our incremental sets g_k are marked in Figure 3.7(b) with different colors: The black pie in corresponds to the common set C, the red area corresponds to g_0, the blue area corresponds to g_1, and the green area corresponds to g_2. Our unique sets γ_k are marked in Figure 3.7(c) as follows: the red area corresponds to γ_0, the blue area corresponds to γ_1, and the green area corresponds to γ_2.

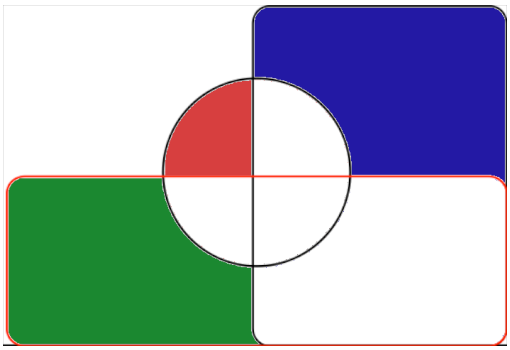
The actual lists of the gene IDs can be found in Appendix C. Finally, we perform a `GO`-term enrichment analysis for each of the order-specific sets γ_k and $g_k, k = 1, \ldots, n$, against the global set G. The biological relevance of the motifs detected using a background model



(a)



(b)



(c)

Figure 3.7: (a) Description of gene sets in the illustration; (b) Incremental construction of order-wise gene sets for GO-term enrichment analysis; (c) Unique construction of order-wise gene sets for GO-term enrichment analysis; see text for details (See page 109).

order is qualitatively assessed by the relevance of the enriched GO-terms to the HIV life cycle and related processes in the host.

3.4 Results and Discussion

3.4.1 Motif Detection

Motifs detected in the HIV integration sequence data by MEME, over a variety of background models (see Table 3.2 for details), are shown in Table 3.4. Specifically, each motif is represented here as the sequence of its highest probability letters at each position. The most detailed description of a motif is its PSPM (see page 85); the PSPMs of all these motifs can be found in appendix A. Column 3 lists $-\log_e$(E-value) for each of the detected motifs, which is a measure of the quality of the motif detected. Lower the E-value (higher the $-\log_e$(E-value)), better the probability that the detected motif is genuine and statistically significant. Column 4 lists the background model used for motif detection.

	Motif	$-\log_e$(E-Value)	Background
1	CCTCAGCCTCCC	1024	
2	AGCTGGGATTACAGGC	1122	
3	CTCCAGCCTGGG	520	MD-0
4	AAACCCCGTCTCTACTAAAAA	1823	
5	CCCCTCCCC	112	
6	AGGCTGAGGCAGGAGG	1085	
7	CGCCTGTACTCCCAGC	957	
8	CCAGCCTGGGCC	562	RP-0
9	CCCCCCAGCCCC	168	
10	GCCACCACGCCCGGCC	515	
11	CCTGTAATCCCAGCTACTCGGG	874	
12	GGCCGGGCGCGGTGGCTCACGC	732	
13	GCCCGGGGGGCAGGGG	068	RP-1
14	GGTTTCACCATGTTGGCCAGGCTGGTCT	1418	
15	CCCAAAGTGCTGGGATTACA	1152	
16	AGCCGGGCGTGGTGGC	445	
17	GCCTCGGCCTCC	393	
18	CAGTGAGCCGAGATCGCGCCACTGCAC	1170	RP-2
19	CCTGTAATCCCAGC	755	
20	TTCTCCTGCCTCAGCCTCCCG	883	
21	GCCTCGGCCTCC	381	
22	CAGCCTCCCAAGTAGCTGGGATT	1123	
23	GCCGGGCGTGGTGGCTCACGCC	1201	RP-3
24	GGCTGGAGTGCAGTGG	493	
25	TGCAGTGAGCCGAGA	217	
26	TAATCCCAGCACTTTGGGAGG	2310	
27	AGCCTGGGCAACATA	976	
28	GGCGTGAGCCACCACGCCCGGC	1979	RP-4
29	AGTGATCCTCCTGCCTCAG	1533	
30	GAGGTTGCAGTGAGCC	280	
31	GGGGCGGGAGGGGCGGGGGCGG	034	
32	CCCAAATTGCTGGGATTACAG	2294	
33	TGGGCAACATAGTGA	777	RP-5
34	AGTGATCCTCCTGCCTCAGCCT	2137	
35	GGCTGGAGTGCAGTGG	1166	

Table 3.4: **List of motifs detected over various backgrounds** Motifs detected by `MEME` over various backgrounds. Column 2: motif, column 3: $-\log_e$(E-value) (lower the E-value (higher the $-\log_e$(E-value) value), better the quality of the motif), column 4: background model (see Table 3.2 for explanation).

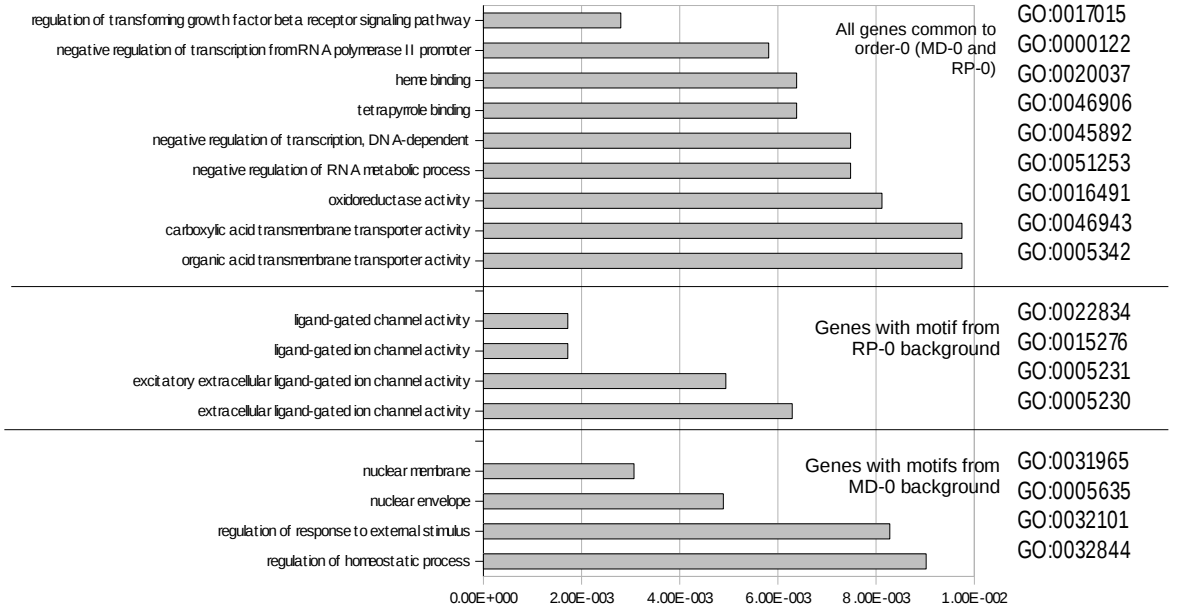


Figure 3.8: Default Background and Markov order-0 Background. The p-values for the enriched GO-terms are plotted with ascending p-values. The enrichments were carried out for genes that contain motifs as detected over MD-0 and RP-0.

3.4.2 Assessment of Biological Relevance

As discussed in Section 3.3.4, we performed GO-term enrichment analysis over sets of genes associated with the motifs detected. There results are presented in Figure 3.8 through Figure 3.10. In these figures the GO-terms related to immune-response are highlighted. Resulting GO-terms (as returned by AmiGO) were filtered using a cutoff of 0.01 on the (hypergeometric) p-value; i.e., only those terms with $p \leq 0.01$ were considered for further analysis. We used the set G (Equation 3.2) as the global set for all our enrichment analyses. Complete listing of our enrichment analysis results is available in Appendix C.

Comparison of RP-0 vs. MD-0 Motifs

As the first step of our GO-term analysis, we compared the biological relevance of motifs detected using the MD-0 and RP-0 backgrounds (Table 3.2). The comparison is motivated by the fact that MD-0 is the MEME default order-0 background model constructed from the HIV integration sequences themselves, and it would be interesting to see if two different background models of the same but low order lead to differing results. This point has been discussed in Section 3.4.3 in greater detail.

Results of this comparison are presented in Figure 3.8.

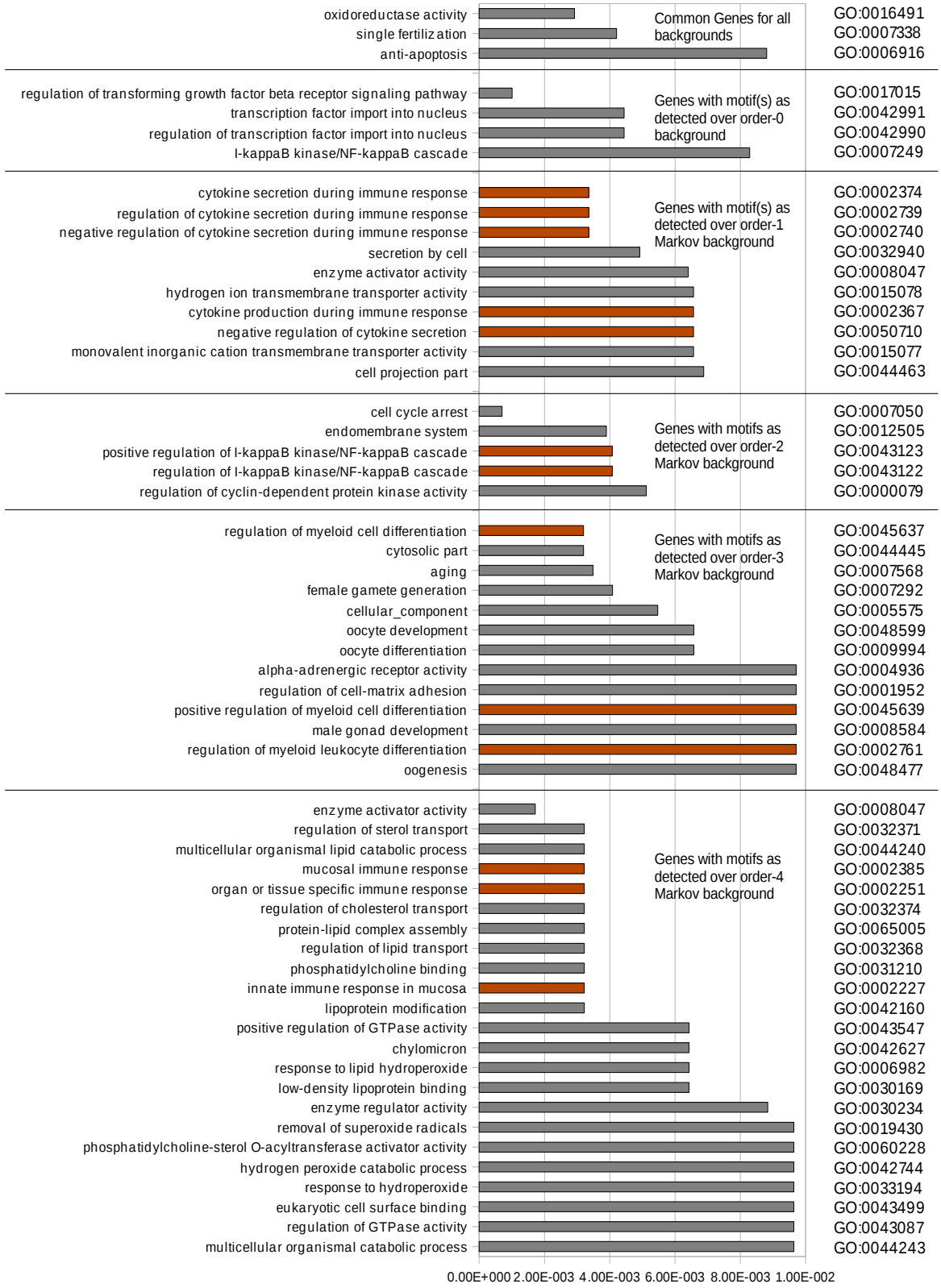


Figure 3.9: In this figure each GO-term enrichment was performed using the sets as mentioned in text (page 110) and Equation 3.3. These enrichments are for the incremental gene lists g_0 through g_4 described earlier.

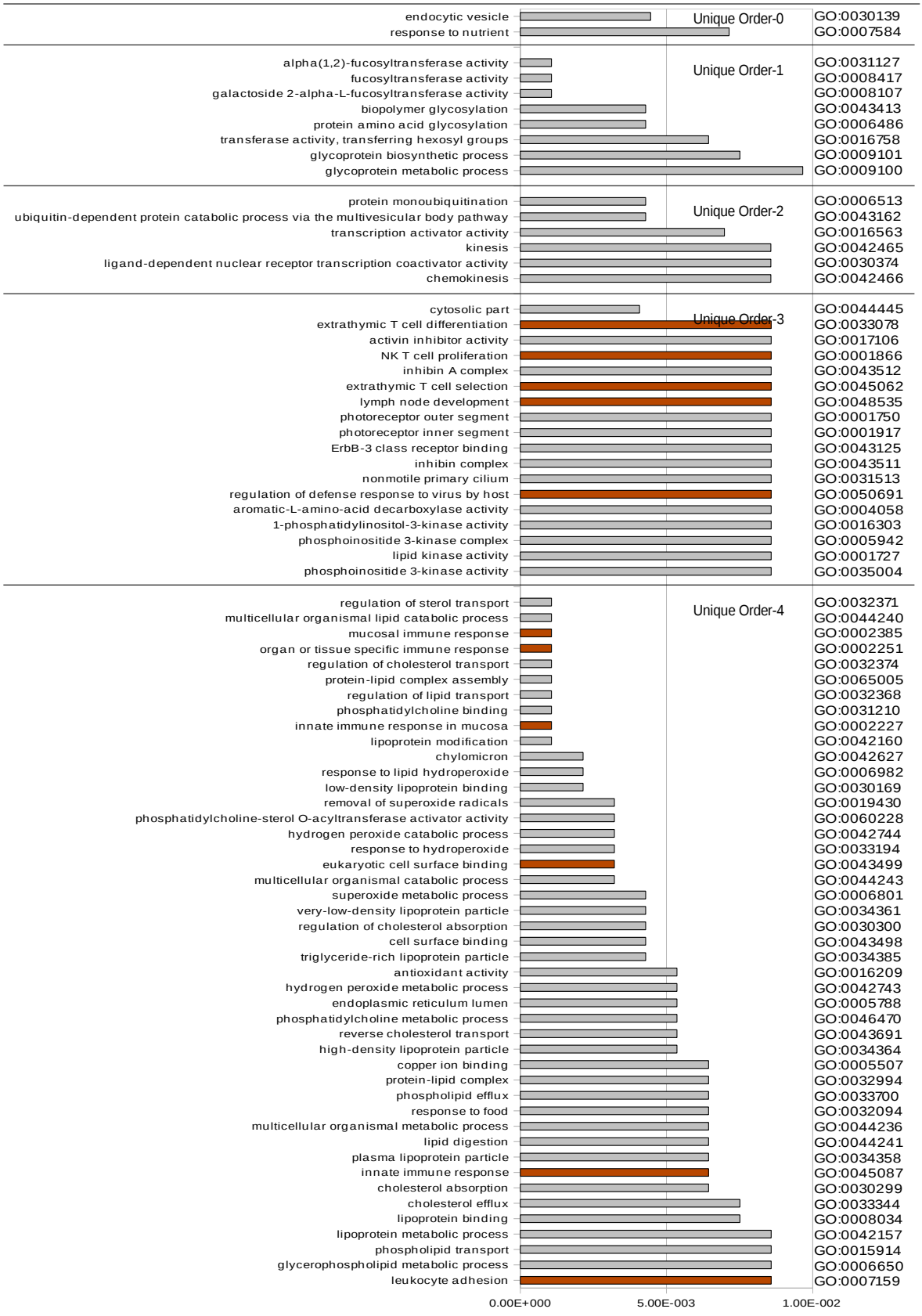


Figure 3.10: In this figure each GO-term enrichment was performed using the sets as mentioned in text (page 111) and Equation 3.4. These enrichments are for unique gene-lists γ_0 through γ_4 described earlier.

Effect of Incremental Gene lists on GO Term Enrichment

In Figure 3.9 we present results for GO-term enrichment for all the backgrounds (RP-0 through RP-5) for increasing orders. These are the results for sets g_0 through g_4 as discussed earlier (Section 3.3.4, equation 3.3). It should be noted here that in this incremental strategy the set g_5 contained only 5 genes and no GO-term enrichment could be done. As described earlier and illustrated in Figure 3.7 the hierarchical and incremental gene-sets were used for GO-term enrichment. All the GO-terms that are potentially related to the HIV lifecycle or the immune response are colored red. It can be clearly seen that as the order of the Markov background increases the motif finds a place in genes that are more related to the immune response. This is expected given the known information about the HIV and copious amounts of literature available (75, 76) also (77, and references therein).

Effect of Unique Gene lists on GO Term Enrichment

In Figure 3.10 we present the data for all the sets γ_0 through γ_4 (see equation 3.4). It can be seen that some more terms are enriched in these gene-sets, over and above those that are illustrated in Figure 3.9, e.g. *NK T cell proliferation*. Thus we can see that as we refine our set of genes under study as a function of the motifs, which in turn are detected over increasing order of the Markov model (background) generated from appropriate sequences the relevance of the genes picked up by the motif to HIV biology increases. This particular observation aptly underlines the need for use of appropriate background for motif detection exercises.

From the number of publications on HIV biology and pathology it is known that HIV perturbs the immune system. So we expected that the GO-Term enrichment should show enrichment of *immune*-related terms such that *immune response*, *mucosal immunity*, *T-cell development* etc. Most interestingly it was seen that as the order of background increased, the motifs detected were found to be occurring with increased frequency in the promoters of genes directly related to immune response.

Thus we can say that the background used for motif detection can directly influence the biological interpretation of such an exercise. Use of appropriate background can lead to observations with interesting insights to the problem.

3.4.3 Discussion

From the results presented in the Figures 3.8 through 3.10, it can be seen that the *enriched* GO-terms change depending on the genes under study. In effect the GO-term enrichment changes according to the background model used for motif detection. This change is apparent even when order-0 Markov models (MD-0 and RP-0) are used as background models. It can be seen that the enriched GO-terms are not identical. Moreover, in the current analysis the G was used super-set. We also repeated these analyses using all the genes as found in the GPD. The results did not change significantly. The explanation for the change in GO-term enrichment could be as follows. In the incremental gene-set g_k (see page 109) we are actually looking at genes that may also contain motifs detected over other higher order backgrounds. This will lead to some *noise* in the enriched terms. However when we use non-incremental unique gene-lists γ_k (see page 109) we have only those genes that specifically have only those motifs as detected over a specific order of background k. These gene-lists are usually smaller than respective incremental gene-lists. Thus, the enrichment of more specific GO-terms could be a result of having a very small number of genes of interest being considered for GO-term enrichment. For the same reason we could not have many genes for analysis that had motifs detected over RP-5 (see Table 3.2 for explanation).

Limitations The results presented in this chapter clearly bring out and underline the importance of choosing appropriate/proper background for motif detection exercises. We would also like to state that these motifs were not checked for their stability or robustness by obtaining the motifs under given condition large (100 times or more) number of times, as described by Thijs *et al.* (13). Such an exercise with the methodology used in the study i.e. the MEME will require an enormous computation power and a very large amount of time. Similarly we have not seen the effect of higher (higher than Order k=5) order background model constructed from the input data on motif detection. Though it has been reported in earlier studies on prokaryotic genomes, that the performance and efficiency of the motif-finding algorithm decreases drastically when order-3 Markov models constructed from the input sequences (13) are used as background model, it needs to be confirmed that the same is true with eukaryotic genomes. We would also like to note in the passing that to critically

assess the role of background on motif detection a few more experiments should be done with following combinations, viz., a) Motif detection on randomly picked genomic sequences using various orders of background models, b) Motif detection in HIV integration sites using various orders of background models built from the same (input sequence) sequences. We believe that such a study will establish, underline and bring out the importance of the work presented in this chapter.

Furthermore, we would also like to add that the GO-term analysis is *indicative* of the biological relevance of the motifs. The results can be further qualified and confirmed using the data available e.g., in the GEO database. It should be possible to analyze the microarray data as obtained from HIV infected cells/patients and do a similar GO-term enrichment study to see how many GO-Terms actually overlap with study presented in this chapter.

3.5 Conclusion

Form all the data presented in this chapter we would like to draw attention to following points,

1. In a motif detection exercise, background should be chosen according to the question being addressed.
2. The choice of the background affects the biological relevance of the results in a motif detection exercise.

References

[1] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–44, 2005.

[2] M. L. Maeder, B. J. Polansky, B. E. Robson, and D. A. Eastman. Phylogenetic Footprinting Analysis in the Upstream Regulatory Regions of the Drosophila Enhancer of split Genes. *Genetics*, 177(3):1377–94, 2007.

[3] P. P. Kumar, S. Mehta, P. K. Purbey, D. Notani, R. S. Jayani, H. J. Purohit, D. V. Raje, D. S. Ravi, R. R. Bhonde, D. Mitra, and S. Galande. SATB1-binding sequences and Alu-like motifs define a unique chromatin context in the vicinity of HIV-1 integration sites. *J Virol*, 2007.

[4] C. Tuerk, S. MacDougal, and L. Gold. RNA pseudoknots that inhibit human immunodeficiency virus type 1 reverse transcriptase. *Proc Natl Acad Sci U S A*, 89(15):6988–92, 1992.

[5] H. Garrido-Hernandez, K. D. Moon, R. L. Geahlen, and R. F. Borch. Design and synthesis of phosphotyrosine peptidomimetic prodrugs. *J Med Chem*, 49(11):3368–76, 2006.

[6] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–80, 1994.

[7] D. G. Higgins, J. D. Thompson, and T. J. Gibson. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol*, 266:383–402, 1996.

[8] Larry Wasserman. *All of Statistics*. Springer Texts in Statistics. Springer, 1 edition, 2004.

[9] Robert V. Hogg, Deceased Allen Craig, and Joseph W. McKean. *Introduction to Mathematical Statistics*. Pearson, 6 edition, 2005.

[10] Arthur Dempster, Nan Laird, and Donald Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39(1):138, 1977.

[11] Jun S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. 1 edition, 2001.

[12] Timothy Lawrence Bailey. *Discovering motifs in DNA and protein sequences:The approximate common substring problem.* PhD thesis, UNIVERSITY OF CALIFORNIA, SAN DIEGO, 1995.

[13] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17(12):1113–22, 2001.

[14] A. V. Favorov, M. S. Gelfand, A. V. Gerasimova, D. A. Ravcheev, A. A. Mironov, and V. J. Makeev. A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*, 21(10):2240–5, 2005.

[15] K. Hiebenthal-Millow, T. C. Greenough, D. B. Bretttler, M. Schindler, S. Wildum, J. L. Sullivan, and F. Kirchhoff. Alterations in HIV-1 LTR promoter activity during AIDS progression. *Virology*, 317(1):109–18, 2003.

[16] M. Dehnert, R. Plaumann, W. E. Helm, and M. T. Hutt. Genome phylogeny based on short-range correlations in DNA sequences. *J Comput Biol*, 12(5):545–53, 2005.

[17] S. Galande, P. K. Purbey, D. Notani, and P. P. Kumar. The third dimension of gene regulation: organization of dynamic chromatin loopscape by SATB1. *Curr Opin Genet Dev*, 17(5):408–14, 2007.

[18] Richard Durbin, Sean R. Eddy, and Anders Krogh. *Biological Sequence Analysis.* Cambridge University Press, 1998.

[19] G. Reinert, S. Schbath, and M. S. Waterman. Probabilistic and statistical properties of words: an overview. *J Comput Biol*, 7(1-2):1–46, 2000.

[20] N. Goldman. Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. *Nucleic Acids Res*, 21(10):2487–91, 1993.

[21] J. van Helden. Regulatory sequence analysis tools. *Nucleic Acids Res*, 31(13):3593–6, 2003.

[22] H. J. Jeffrey. Chaos game representation of gene structure. *Nucleic Acids Res*, 18(8):2163–70, 1990.

[23] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol*, 16(10):1391–9, 1999.

[24] P. Deschavanne and P. Tuffery. Exploring an alignment free approach for protein classification and structural class prediction. *Biochimie*, 90(4):615–25, 2008.

[25] J. Joseph and R. Sasikumar. Chaos game representation for comparison of whole genomes. *BMC Bioinformatics*, 7:243, 2006.

[26] W. R. Pearson. Using the FASTA program to search protein and DNA sequence databases. *Methods Mol Biol*, 24:307–31, 1994.

[27] J. S. Almeida, J. A. Carrico, A. Maretzek, P. A. Noble, and M. Fletcher. Analysis of genomic sequences by Chaos Game Representation. *Bioinformatics*, 17(5):429–37, 2001.

[28] J. Hudak and M. A. Mcclure. A comparative analysis of computational motif-detection methods. *Pac Symp Biocomput*, pages 138–49, 1999.

[29] S. Henikoff, J. G. Henikoff, W. J. Alford, and S. Pietrokovski. Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene*, 163(2):GC17–26, 1995.

[30] M. A. Saqi and M. J. Sternberg. Identification of sequence motifs from a set of proteins with related function. *Protein Eng*, 7(2):165–71, 1994.

[31] L. Brocchieri and S. Karlin. A symmetric-iterated multiple alignment of protein sequences. *J Mol Biol*, 276(1):249–64, 1998.

[32] R. F. Smith and T. F. Smith. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng*, 5(1):35–41, 1992.

[33] A. F. Neuwald, J. S. Liu, D. J. Lipman, and C. E. Lawrence. Extracting protein alignment models from the sequence database. *Nucleic Acids Res*, 25(9):1665–77, 1997.

[34] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*, 34(Web Server issue):W369–73, 2006.

[35] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol*, 235(5):1501–31, 1994.

[36] R. Hughey and A. Krogh. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput Appl Biosci*, 12(2):95–107, 1996.

[37] J. T. Wang, T. G. Marr, D. Shasha, B. A. Shapiro, and G. W. Chirn. Discovering active motifs in sets of related protein sequences and using them for classification. *Nucleic Acids Res*, 22(14):2769–75, 1994.

[38] J. G. Henikoff and S. Henikoff. Blocks database and its applications. *Methods Enzymol*, 266:88–105, 1996.

[39] J. Jurka, V. V. Kapitonov, O. Kohany, and M. V. Jurka. Repetitive sequences in complex genomes: structure and evolution. *Annu Rev Genomics Hum Genet*, 8:241–59, 2007.

[40] A. J. Pellionisz. The Principle of Recursive Genome Function. *Cerebellum*, 2008.

[41] S. Carteau, C. Hoffmann, and F. Bushman. Chromosome structure and human immunodeficiency virus type 1 cDNA integration: centromeric alphoid repeats are a disfavored target. *J Virol*, 72(5):4005–14, 1998.

[42] V. V. Lunyak. Boundaries. Boundaries...Boundaries??? *Curr Opin Cell Biol*, 20(3):281–7, 2008.

[43] M. G. Schueler and B. A. Sullivan. Structural and functional dynamics of human centromeric chromatin. *Annu Rev Genomics Hum Genet*, 7:301–13, 2006.

[44] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–14, 1993.

[45] S. Aerts, G. Thijs, B. Coessens, M. Staes, Y. Moreau, and B. De Moor. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*, 31(6):1753–64, 2003.

[46] W. Cochran and G Cox. *Experimental Designs*. Wiley, second edition, 1957.

[47] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

[48] E. Eden, D. Lipson, S. Yogev, and Z. Yakhini. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol*, 3(3):e39, 2007.

[49] K. Marchal, G. Thijs, S. De Keersmaecker, P. Monsieurs, B. De Moor, and J. Vanderleyden. Genome-specific higher-order background models to improve motif detection. *Trends Microbiol*, 11(2):61–6, 2003.

[50] J. M. Coffin. Retroviral DNA integration. *Dev Biol Stand*, 76:141–51, 1992.

[51] W. A. Haseltine. Molecular biology of the human immunodeficiency virus type 1. *FASEB J*, 5(10):2349–60, 1991.

[52] J. Lehmann-Che and A. Saib. Early stages of HIV replication: how to hijack cellular functions for a successful infection. *AIDS Rev*, 6(4):199–207, 2004.

[53] M. Seiki, R. Eddy, T. B. Shows, and M. Yoshida. Nonspecific integration of the HTLV provirus genome into adult T-cell leukaemia cells. *Nature*, 309(5969):640–2, 1984.

[54] A. R. Schroder, P. Shinn, H. Chen, C. Berry, J. R. Ecker, and F. Bushman. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*, 110(4):521–9, 2002.

[55] H. Zhou, G. J. Rainey, S. K. Wong, and J. M. Coffin. Substrate sequence selection by retroviral integrase. *J Virol*, 75(3):1359–70, 2001.

[56] X. Wu, Y. Li, B. Crise, and S. M. Burgess. Transcription start regions in the human genome are favored targets for MLV integration. *Science*, 300(5626):1749–51, 2003.

[57] R. S. Mitchell, B. F. Beitzel, A. R. Schroder, P. Shinn, H. Chen, C. C. Berry, J. R. Ecker, and F. D. Bushman. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol*, 2(8):E234, 2004.

[58] X. Wu, Y. Li, B. Crise, S. M. Burgess, and D. J. Munroe. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J Virol*, 79(8):5211–4, 2005.

[59] J. F. Hughes and J. M. Coffin. Human endogenous retroviral elements as indicators of ectopic recombination events in the primate genome. *Genetics*, 171(3):1183–94, 2005.

[60] T. L. Bailey and M. Gribskov. Methods and statistics for combining motif match scores. *J Comput Biol*, 5(2):211–21, 1998.

[61] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9, 2000.

[62] O. Kohany, A. J. Gentles, L. Hankus, and J. Jurka. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics*, 7:474, 2006.

[63] T. L. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol*, 3:21–9, 1995.

[64] Y. Ponty, M. Termier, and A. Denise. GenRGenS: software for generating random genomic sequences and structures. *Bioinformatics*, 22(12):1534–5, 2006.

[65] Makato Matsumoto and Takuji Nishimura. Mersenne Twister: A 623-dimensionally equidistributed uniform pseudorandom generator. *ACM Transactions on Modelling Computer Simulation*, (8):3–30, 1998.

[66] W. N. Grundy, T. L. Bailey, C. P. Elkan, and M. E. Baker. Meta-MEME: motif-based hidden Markov models of protein families. *Comput Appl Biosci*, 13(4):397–406, 1997.

[67] T. Tsukahara, H. Agawa, S. Matsumoto, M. Matsuda, S. Ueno, Y. Yamashita, K. Yamada, N. Tanaka, K. Kojima, and T. Takeshita. Murine leukemia virus vector integration favors promoter regions and regional hot spots in a human T-cell line. *Biochem Biophys Res Commun*, 345(3):1099–107, 2006.

[68] M. Mistry and P. Pavlidis. Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, 9:327, 2008.

[69] E. Dimmer, T. Z. Berardini, D. Barrell, and E. Camon. Methods for gene ontology annotation. *Methods Mol Biol*, 406:495–520, 2007.

[70] V. Beisvag, F. K. Junge, H. Bergum, L. Jolsum, S. Lydersen, C. C. Gunther, H. Ramampiaro, M. Langaas, A. K. Sandvik, and A. Laegreid. GeneTools–application for functional annotation and statistical hypothesis testing. *BMC Bioinformatics*, 7:470, 2006.

[71] Robert V. Hogg and Allen T. Craig. *Introduction to Mathematical Statistics*. Prentice Hall, 1995.

[72] The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*, 34(Database issue):D322–6, 2006.

[73] S. Falcon and R. Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–8, 2007.

[74] A. W. F. Edwards. *Cogwheels of the Mind: the story of Venn diagrams*. Johns Hopkins University Press, 2004.

[75] M. Alfano, A. Crotti, E. Vicenzi, and G. Poli. New players in cytokine control of HIV infection. *Curr HIV/AIDS Rep*, 5(1):27–32, 2008.

[76] A. L. Abujamra, R. A. Spanjaard, I. Akinsheye, X. Zhao, D. V. Faller, and S. K. Ghosh. Leukemia virus long terminal repeat activates NFkappaB pathway by a TLR3-dependent mechanism. *Virology*, 345(2):390–403, 2006.

[77] F. Bushman. Targeting retroviral integration? *Mol Ther*, 6(5):570–1, 2002.

Chapter 4

Deciphering Gene Regulatory Networks

4.1 Introduction

The problem of deciphering tissue-specificity signature(s) in a promoter sequence is of considerable interest from many perspectives. For instance, based on diverse evidence, it is believed that primary signatures of tissue-specificity are hidden in the promoter sequence itself and are responsible for regulating gene expression in a tissue-specific manner. From a systems biology perspective, knowledge of tissue-specificity signatures in promoter sequences could help in inferring gene regulatory networks that control tissue-specific gene expression. Mutations in promoters as a cause of carcinogenesis, for example, highlight the "applied" perspective on signatures of tissue-specificity hidden in a promoter sequence such that it would help in designing an appropriate strategy for gene therapy like approaches to treat cancer.

In this chapter we propose basic premise and methodology to identify characteristic primary sequence features of human promoter sequences that give the corresponding gene a tissue-specific expression profile. In particular, this chapter proposes to focus on promoters of genes that are transcribed by RNA polymerase II. Various computational approaches based on classification and clustering methodologies may be employed to address this problem. These approaches are described in this chapter.

The methodologies proposed to be developed to predict tissue-specific expression of gene(s) will help plan better experiments by reducing the amount of trial-and-error. This is of great practical value given the escalating costs of molecular biology research today. Furthermore, the ability to predict pattern of expression of gene based solely on the primary sequence pat-

terns will also help in designing useful vectors for the still-under-development gene therapy.

4.2 Genesis of the Problem

Promoters are short segments of genomic DNA located immediately adjacent to the transcriptional start sites (TSS) of genes. They are recognized by both general and sequence-specific transcription factors during transcription initiation, and serve to integrate signals from multiple cellular pathways to promote stringently regulated and specific expression of gene(s). A large complex protein structure referred to as the pre-initiation complex is assembled on all active promoters. Presence of the pre-initiation complex is a hallmark of promoters and this feature distinguishes them from bulk of genome. Many promoters have been identified based on in vitro and in vivo experiments using cell line and/or animal models. Knowledge of promoter sequences is essential for understanding the mechanisms of gene regulation during development and differentiation. A typical eukaryotic promoter consists of a complex array of cis-regulatory elements (sequence motifs) (1, and references therein), such as TATA box, Initiator (INR), TFIIB Recognition Element (BRE), downstream promoter element (DPE), downstream core element (DCE), motif 10 element (MTE), etc. Additionally, there are other signals superimposed in the promoter (the upstream cis-regulatory regions) regions of the genes such as the nucleosome positioning (2).

Computational analysis of genomic sequences has traditionally focused on computational identification of promoters. However, not much work has been done with respect to predicting tissue- specificity of known and predicted promoters. In this chapter methodologies to address this particular problem using a systems biology approach are discussed.

4.2.1 Rationale of the Study

The problem of deciphering tissue-specificity of gene expression is of considerable interest from multiple perspectives. From a fundamental science perspective, tissue-specific gene expression is believed to be controlled by transcription factors (3). Further analysis of gene expression from integrated promoter and transgenic mice studies suggests that the promoter alone is enough to drive transcription in a tissue-specific manner. This suggests that signatures of the tissue-specificity could be hidden in the promoter sequence itself.

These primary sequence features could regulate tissue-specific gene expression in three possible ways: via organizing the gene into a unique chromatin context, or by providing sites for binding of a set of transcription factors, or both. Chromatin is a complex of DNA and interacting proteins (including the structural proteins such as the histones) and various transcription factors. Often it is observed that the specificity of expression is derived from the combinatorial effect of multitude of factors, and hence analyzing such global networks of factors seems to be the next challenge in the field of regulation of gene expression.

From a systems biology perspective, such signatures could help in inferring gene regulatory networks that control tissue-specific gene expression. Such attempts have been made in single-cell eukaryotes such as the yeast *Saccharomyces cereviseae* but not for higher eukaryotes. These attempts have also revealed putative novel interactions and relationships which could be subsequently tested in the lab and verified. Further, even a probabilistic assignment of a promoter sequence to a particular tissue will greatly help plan better experiments by reducing the trial-and-error in determining expression patterns of the corresponding gene(s).

From a purely applied perspective, many studies such as described in Harada *et al.* (4). Harada *et al.* have demonstrated that carcinogenesis is positively linked with mutations in promoters of known tissue-specific genes (4). Furthermore, tissue-specificity of a promoter has also been utilized to construct gene therapy vectors for treatment of cancer. Another highly applied perspective is offered by J. Adjaye, which we simply quote here (5):

> *The elucidation, unraveling and understanding of the molecular basis of transcriptional control during preimplantation development is of utmost importance if we are to intervene and eliminate or reduce abnormalities associated with growth, disease and infertility...*

In this chapter methodologies to identify characteristic primary sequence features of human promoter sequences that give the corresponding gene(s) a tissue-specific expression profile are discussed. The gene expression in higher organisms that have multiple terminally differentiated tissues involves an added level of regulation. The mechanisms that control the tissue-specific expression of genes are largely unknown. In this chapter methodologies to this end are described.

4.2.2 Hypothesis

Usually in the literature available today a promoter is considered to be tissue-specific if the expression profile of the corresponding gene is tissue-specific. In accordance with `GeneCards` (6) and `GeneNote` (7), for the purpose of this chapter we define tissue-specific expression as follows: a gene is considered tissue-specific if, in the analyzed microarray data, it is at least two-fold over-expressed in given tissue and is either not detectable or normally expressed in all other tissues. In particular, this chapter will focus on promoters of genes that are transcribed by PolII (DNA dependent RNA polymerase II). Such promoters are usually positioned somewhere around -300 to -50 bp upstream of the Transcription Start Site (TSS). However, for our analysis we may also consider regions up to 1000 bp upstream of the TSS, so as to cover and account for a remote possibility that the core promoter is situated upstream of the generally expected position.

It is further hypothesized that the chromatin context of each promoter is unique and may be responsible for driving the expression profile of the gene controlled by it. It is further hypothesized that the information for such unique higher order chromatin structure may be hidden within the primary sequence of the promoter itself. Thus it may be possible to uncover these hidden signatures using purely computational means guided by biological insight.

These signatures, in conjunction with tissue-specific transcription factors or their combinations may dictate tissue-specific gene expression. Thus, a systems biology approach is required to understand this complex interplay and the networks between various cis and trans-acting elements. The human genome is composed of a very small fraction of protein coding sequence (upto 2%) while the remainder bulk consists of non-protein coding sequence. Within the 2% of protein coding sequence, a very tiny fraction corresponds to the promoter elements. Thus, promoter sequence could be considered as highly specialized DNA sequences, and therefore we expect that they could be classified according to their tissue-specificity. Once such classifiers/predictors are constructed and refined, then we could use them to predict functional attributes of any given promoter sequence.

4.3 Background

Methodologies to determine whether a given DNA sequence may contain a promoter or is itself likely to be a promoter are readily available. Multiple studies have attempted to classify promoter sequences based on experimental data (8). A study by Frech, Quandt and Werner (9) investigated tissue-specificity signatures in muscle actin promoters across many species. Recent studies also demonstrate definitively that the cis-regulatory elements indeed control the tissue-specific expression of genes (10). A novel approach by Schug *et al.* (11) illustrates that *Shannon entropy* measure on microarray data can be used to rank genes according to their tissue-specificity.

Many recent studies have been geared towards finding features of promoter sequences that may be associated with tissue-specificity of gene expression. For example, Fitzgerald *et al.* (12) have demonstrated clustering of specific oligomers close to the Transcription Start Site (TSS) in human promoters. Smith *et al.* (13) have demonstrated that the proximal promoter and cis-acting elements that control tissue-specific transcription in the mouse and the human. Zhang *et al.* (14) have demonstrated that clustering of degenerate transcription factor binding site motifs on a promoter is a general feature of mammalian genome. Xie *et al.* (15, 16) have created a catalog of regulatory motifs in human promoters. They have also used known functional cis-regulatory modules (CRM)s in the proximal promoters to predict tissue-specific gene expression. Their results indicate that information in the proximal promoter can be used to predict differential expression of downstream target transcripts in terminally differentiated human and mouse tissues with significant accuracy. Schug *et al.* (11) have illustrated the role of CpG islands , the TATA box, YY1 and SP1 recognition sites on a promoter and the expression of corresponding gene(s). Tools such as `TRANSFAC` (17, 18) and `MATCH`$^{\text{TM}}$ predict the presence of transcription factor recognition sites on a given sequence (for review see (19) and (20)).

4.3.1 Relevance and Expected Output

There are several grey areas in the current state-of-the-art methodologies and tools for promoter classification according to tissue-specificity. For example Smith *et al.* (13) have shown that it is indeed possible to construct tissue-wise predictors, however their efficacy and effi-

ciency are questionable, especially in the light of possible input of *housekeeping* genes. The housekeeping genes are defined as follows:

housekeeping genes Genes that are always expressed (i.e. they are said to be constitutively expressed) due to their constant requirement by the cell.

Furthermore, most of the studies on the subject of tissue-specific promoter classification are actually concerned with finding motifs, transcription factor binding sites etc. Although this is of considerable biological relevance and value, the evaluation of existing classification and clustering methodologies for the purpose of tissue-wise promoter classification, and perhaps development of new biologically-motivated ones, is still an emerging and open area of research.

4.4 Exploratory Data Analysis

The exploratory data analysis was carried out in 3 ways, viz., analysis of oligomer frequencies, analysis of the transcription factor networks, and study of gene networks.

4.4.1 Studies Using Distance Measures

We used promoters of the known tissue-specifically expressed genes in this study. Briefly, a normalized compression distance was defined (see Appendix B for details of formulation of compression based distance measures and their use in classification and clustering of DNA sequences). For the purpose of this study we used the sequence data as described in Smith *et al.* (13). These sequences are essentially regions around the transcription start site, from -1000 to +100 with respect to the TSS. In the supplementary data of the said paper, these sequences are available in the `FASTA` format and classified according to the tissue. We used these sequences as inputs for the compression distance calculating algorithm. We generated a distance matrix of all promoters using the compression based distances. Briefly, we obtained promoter sequences from the `TCAT` database as mentioned in Smith *et al.* (13). These promoters are classified as specific to a tissue based on microarray analysis and other statistical parameters as described earlier by Smith *et al.* (13). As described and defined in Appendix B, we used the symmetrized normalized compression distanceas a metric of choice to distinguish

between the promoters of genes that expressed in tissue-specific manner. On such obtained distance matrices we used the various hierarchical clustering techniques. However, these techniques failed to distinguish the promoter sequence of tissue-specifically expressed genes from each other. To the best of our knowledge use of compression distance to distinguish promoters that drive tissue-specific expression of genes has not been attempted. There have been attempts of using mutual information, *Shannon's Entropy* and other information complexity measures to classify DNA sequences. However, such methodologies have failed to address the problem adequately. Other compression based distance measures were used, and found not to be adequate to distinguish promoters of tissue-specifically expressed genes (based on tissues).

In addition to the compression based distance measure, we also tried using the *Levenshtein Distance* as a metric to distinguish the promoters of tissue-specifically expressed genes from one other. In information theory and computer science, the Levenshtein distance is a metric for measuring the amount of difference between two sequences. It is also known as the *Edit distance*. There are following bounds on the Levenshtein distance (as obtained from Wikipedia[1]) viz.,

- It is always at least the difference of the sizes of the two strings
- It is at most the length of the longer string
- It is zero if and only if the strings are identical
- If the strings are of same size, the *Hamming Distance* is an upper bound on the Levenshtein Distance.

This metric also failed to distinguish between promoters of the various tissue-specifically expressed genes. The reason(s) why compression based distance measures were able to classify artificial and natural DNA sequences is not entirely clear. This measure has been used successfully in authorship attribution. The probable reason for the effectiveness of the compression based distance measure(s) is in basis of information theory and computer science. The compression algorithms were able to 'sense' the information in the genomic DNA sequences, whereas in the artificial sequences such information was absent. However, it cannot

[1] `http://www.wikipedia.org`

be stated in any definite manner 'what' this information is. In the same line it is possible that the compression algorithms are not able to decipher the tissue-specific expression information from the promoter sequences.

4.4.2 Oligomer Frequency Analysis

Initially we obtained list of tissue-specifically expressed genes from `GeneNote` and `GeneCard` databases. The lists were basically available as gene-symbol(s) (Gene symbols are the officially designated short representations of the gene-names as standardized by the HUGO, these symbols are used in most of the standard databases that store the gene information). Using these gene symbols and the `CCDS` database (Comprehensive cDNA Database), the exact genomic locations of the genes were obtained. Sometimes there are multiple entries for each gene, so the first entry was used to obtain information, other entries were ignored. Similarly, for many genes multiple transcription start sites have been documented. For the purposes of these studies, the first reported transcription start site was used. All the information that was ignored during the study may be biologically relevant, but no reasonable estimate can be made about such ignored information for the lack of literature. Furthermore, detailed analyses have to be carried out to obtain a complete or near-complete picture of the upstream regulatory regions of the genes.

Using the obtained genomic locations, we downloaded the upstream regulatory regions. In house scripts and well documented NCBI APIs were used to obtain such sequences. We defined, for the purposes of these studies, the upstream regulatory region to be from -2000 to -1 with respect to the transcription start site. The rationale being, core promoter will surely be represented in these regions along with most important proximal cis-regulatory region(s). The sequences were obtained in `FASTA` format. As the sequences were downloaded from the NCBI `RefSeq` database, the downloaded sequences did not have the information regarding the gene the sequences were related to. To simplify further analyses, we added the Gene Accession Number[2] to the description line of each `FASTA` sequence downloaded.

These sequences were then analyzed for their DNA oligomer frequency counts. Briefly,

[2]Gene Accession Number is a unique alpha-numeric identifier of each entry in the gene database. There are specific conventions followed in these nomenclatures and the complete list of these conventions can be found the help documents at the NCBI.

for each sequence, all possible occurrences for all possible tetramers were generated. The count of each possible tetramer for each sequence was then divided by number of possible tetramers in the given sequence, and a normalized tetramer frequency was generated. The normalized tetramer counts were generated as follows.,

$$N_c = \frac{C_{oligomer}}{N - l + 1} \tag{4.1}$$

N_c is the normalized count of the the oligomer, N is the length of the sequence, l is the length of the oligomer, and $C_{oligomer}$ is the count of the given oligomer in the given sequence.

This particular measure will generate counts of oligomers (tetramers in this case) such that the numbers will be comparable even if the length of sequences under study vary. These normalized frequencies are plotted here. In Figures 4.1 through 4.4 the normalized frequencies of the tetramers are plotted. In all the graphs the tetramer index on the X-axis is from `AAAA` to `TTTT` (0 to 255, total 256). It can be seen clearly that the frequencies of the the tetramers show nearly identical patterns across the different tissues. We believe this is because of the nature of sequences from which these frequencies were generated. As discussed earlier these graphs were generated from the promoter sequences. That could be the reason for have similar pattern of tetramer frequencies in the sequences. However, when normalized tetramer counts for two or more tissues are plotted together, certain differences become apparent.

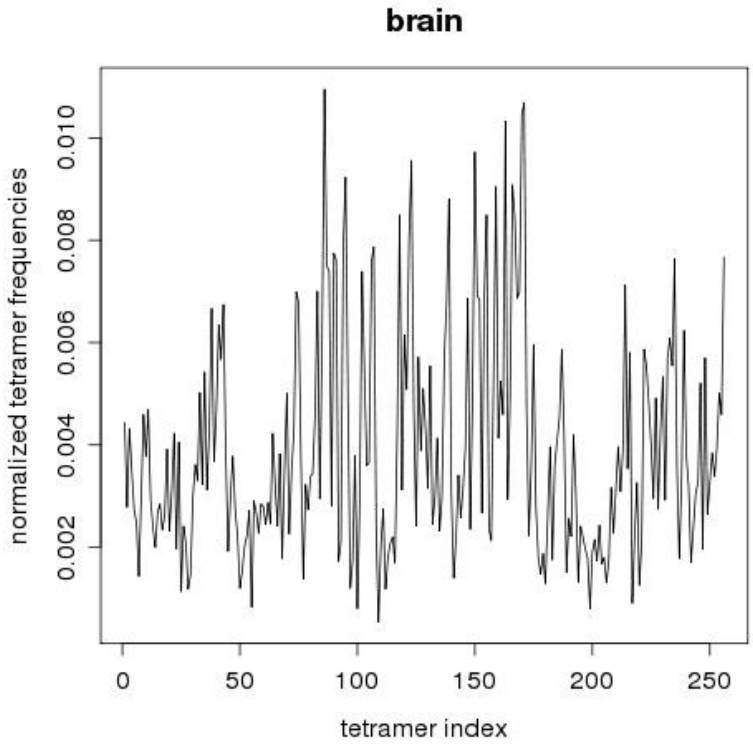


Figure 4.1: Brain–normalized tetramer frequencies for promoters obtained from the `TCAT` database (13) are plotted on the Y-axis and the tetramer-index on the X-axis.

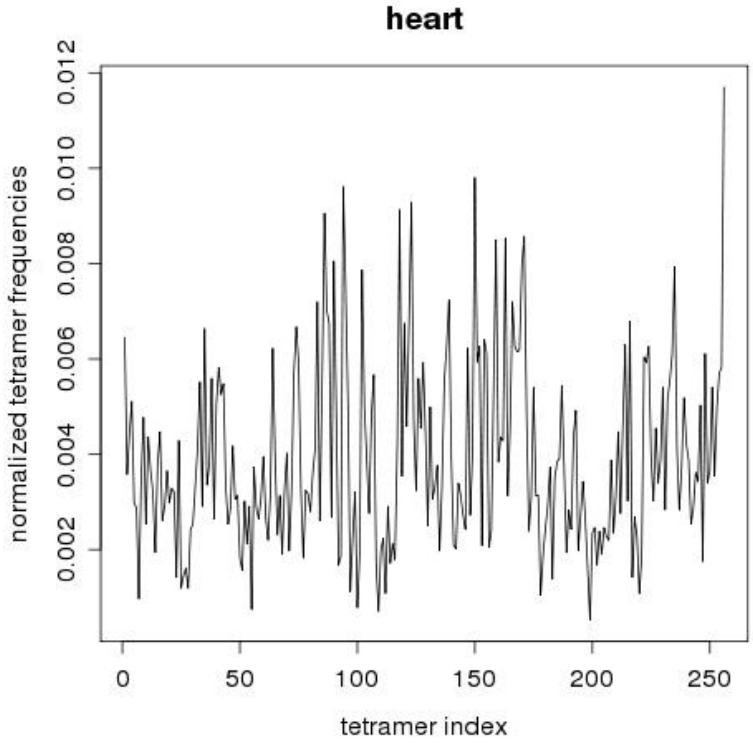


Figure 4.2: Heard–normalized tetramer frequencies for promoters obtained from the `TCAT` database (13) are plotted on the Y-axis and the tetramer-index on the X-axis.

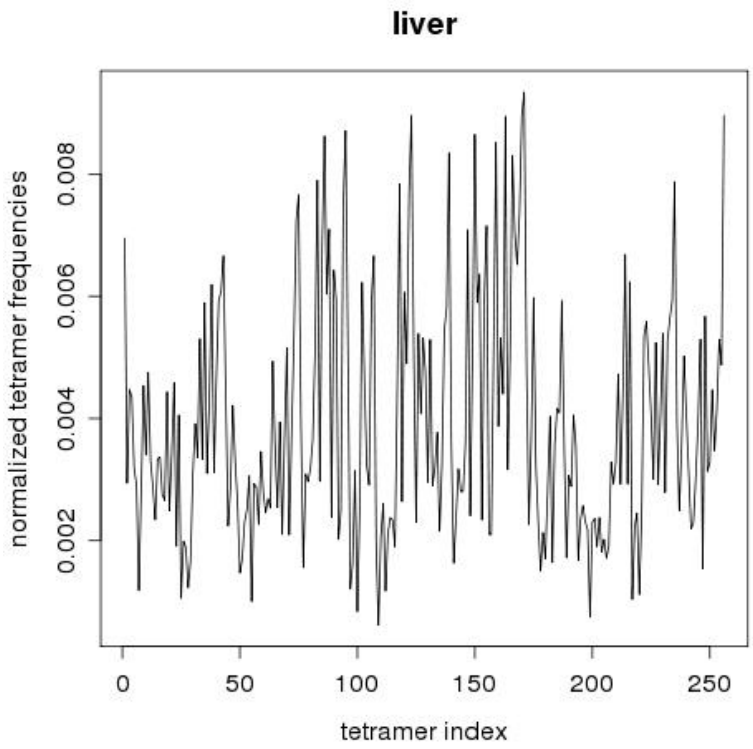


Figure 4.3: Liver–normalized tetramer frequencies for promoters obtained from the TCAT database (13) are plotted on the Y-axis and the tetramer-index on the X-axis.

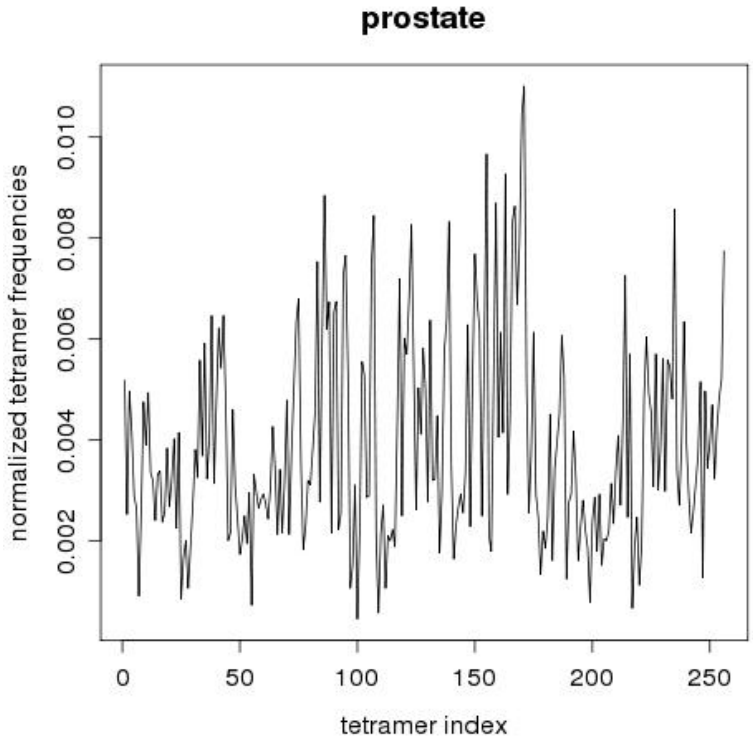


Figure 4.4: Prostate–normalized tetramer frequencies for promoters obtained from the TCAT database (13) are plotted on the Y-axis and the tetramer-index on the X-axis.

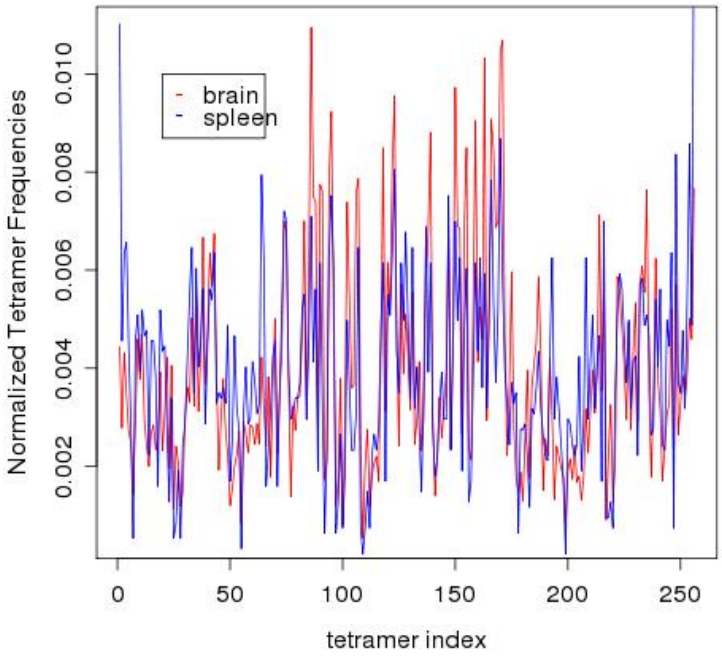


Figure 4.5: Normalized frequencies show distributions unique to the tissue-types. Tissues with dissimilar developmental lineage have widely varying distribution of normalized tetramer frequencies

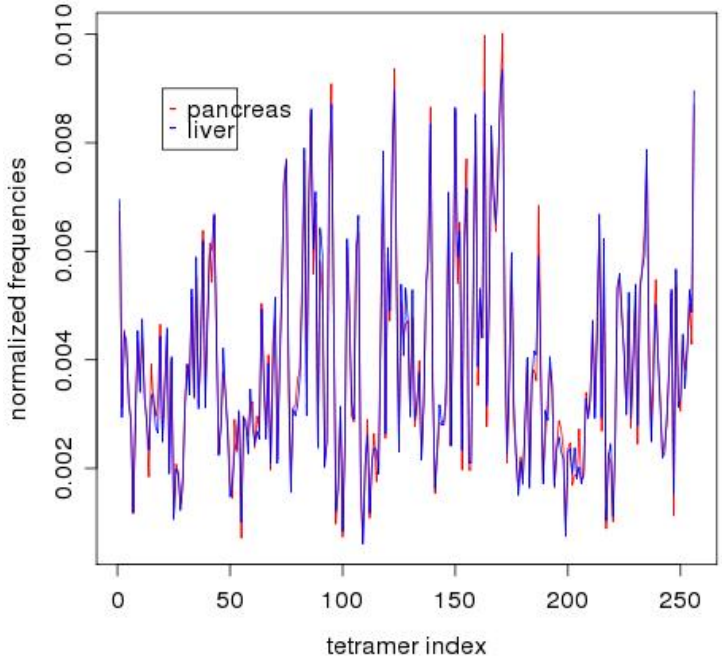


Figure 4.6: Tissues with similar developmental lineage show similar normalized frequency.

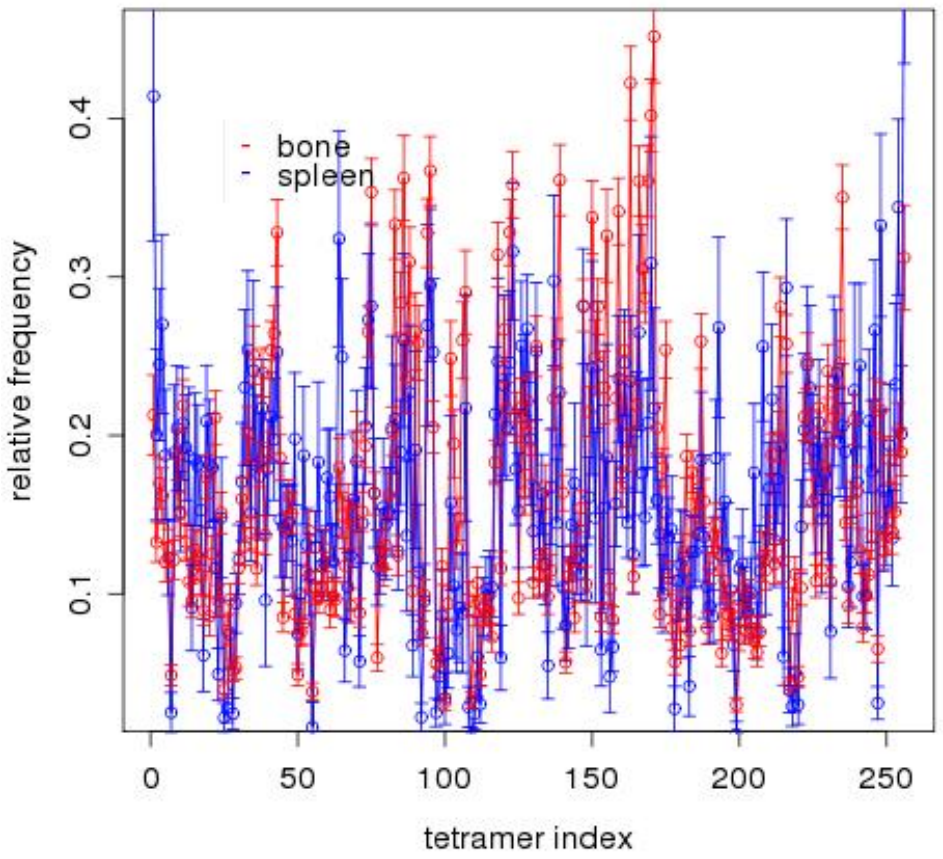


Figure 4.7: Relative tetramer frequencies in bone and spleen. The relative tetramer frequencies (tetramer frequencies normalized by frequency of the maximum occurring tetramer) plotted with error-bars for each tetramer. At least for few tetramers the difference in the relative frequency is statistically significant.

Moreover, when such frequencies are plotted for similar tissues, e.g. tissues with similar developmental lineage, such as liver and pancreas, majority of these differences disappear as seen clearly in Figure 4.6. It however remains to be seen if these differences are systematic. However, even if these differences are not statistically significant at the level of normalized tetramers, they may become statistically significant when such oligomer frequencies are studied and analyzed for oligomers larger than 4-mers.

In addition we *normalized* frequencies we also looked at the *Relative* tetramer frequencies. These are frequencies normalized with frequency of the maximum occurring tetramer. However, when the relative frequencies are plotted with error bars (standard error about the mean as calculated for each oligomer for all the sequences assigned to a single tissue), the differences for frequencies of at least a few tetramers show up as statistically significant. The difference is especially stark when compared across tissues which are functionally and physiologically very distinct e.g. bone and spleen (Figure 4.7).

We therefore conclude that the information about the regulation of gene expression is in the promoter sequences. Moreover, the information about the function of a given DNA sequence e.g. as a promoter, is also in the primary sequence. Simple analysis of these sequences, in terms of their composition can be a good indication of this phenomenon. We believe that there is a lot of potential for various emerging and established statistical theories for such problems. As this problem ultimately boils down to analyzing 'strings' of DNA, many techniques from the *Language theory* can be applied here. These approaches are indeed being considered and have shown some encouraging results (21).

4.4.3 Transcription Factor Co-occurrence Network

As part of exploratory data analysis, and to derive proof-of-concept we obtained sequence data by integrating information from various sources. We processed this data in tissue-wise manner, through `MATCH`^{TM}available on the web, to identify transcription factor binding sites on each of the promoter sequences in this data. Using a combination of settings for this tool, we restricted this search to high-quality weight matrices for transcription factors from vertebrates only, and the score cutoffs were chosen to minimize false positives as well as false negatives.

The output from the MATCH^{TM}(22) were parsed and this information was used to compute a heuristic distance for each pair of transcription factors (TF). In essence, this distance measure counts the co-occurrences of a pair of TFs on the same promoter sequence, averaged across all promoters belonging to the same tissue. Next these distances were used to construct tissue-wise TF networks. The link weight (weights of the nodes in this network are the frequencies of occurrence of individual TFs averaged across all promoters belonging to the same tissue) is number of times the given pair of TFs occurs in a given tissue. Finally, the connection-graphs were visualized as TF-networks using a software tool called the graphviz[3] using a color code[4] that is uniquely determined by the node and link weights.

[3]Graphviz (short for Graph Visualization Software) is a package of open source tools initiated by AT&T Research Labs for drawing graphs specified in DOT language scripts. It also provides libraries for software applications to use the tools. Graphviz is free software licensed under the Common Public License (http://www.graphviz.org).

[4]The RGB color space is represented in hexadecimal format, such that #000000 means black and #ffffff means white. Each pair of position gives the amount of R (red) or G (green) or B (blue) color present. So #ff0000 becomes a red color and so on. Additionally each hexadecimal number can also have a binary/decimal representation. Such a color space was divided across the frequency spectrum obtained for the co-occurrence, and the color was assigned such that frequencies of co-occurrence were scaled and represented as hexadecimal numbers that were in turn mapped to the RGB colorspace. Maximum frequency received the red color and the minimum frequency received the blue color.

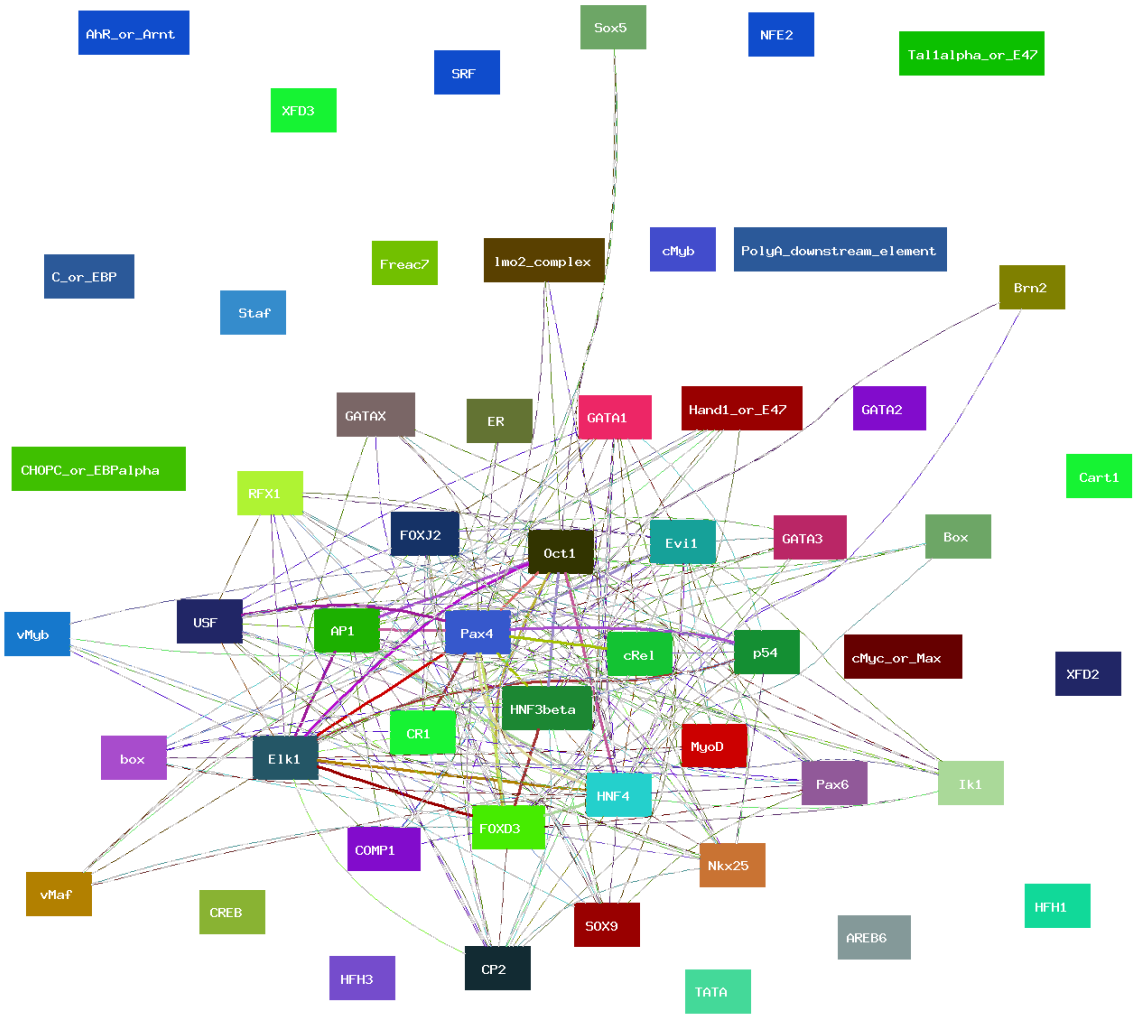


Figure 4.8: CD4^{+}-T cell transcription factor co-occurrence network. The MATCH^{TM} output for the sequences obtained from the TCAT database visualized as transcription-factor co-occurrence network with color coding as mentioned on page 140, footnote 4.

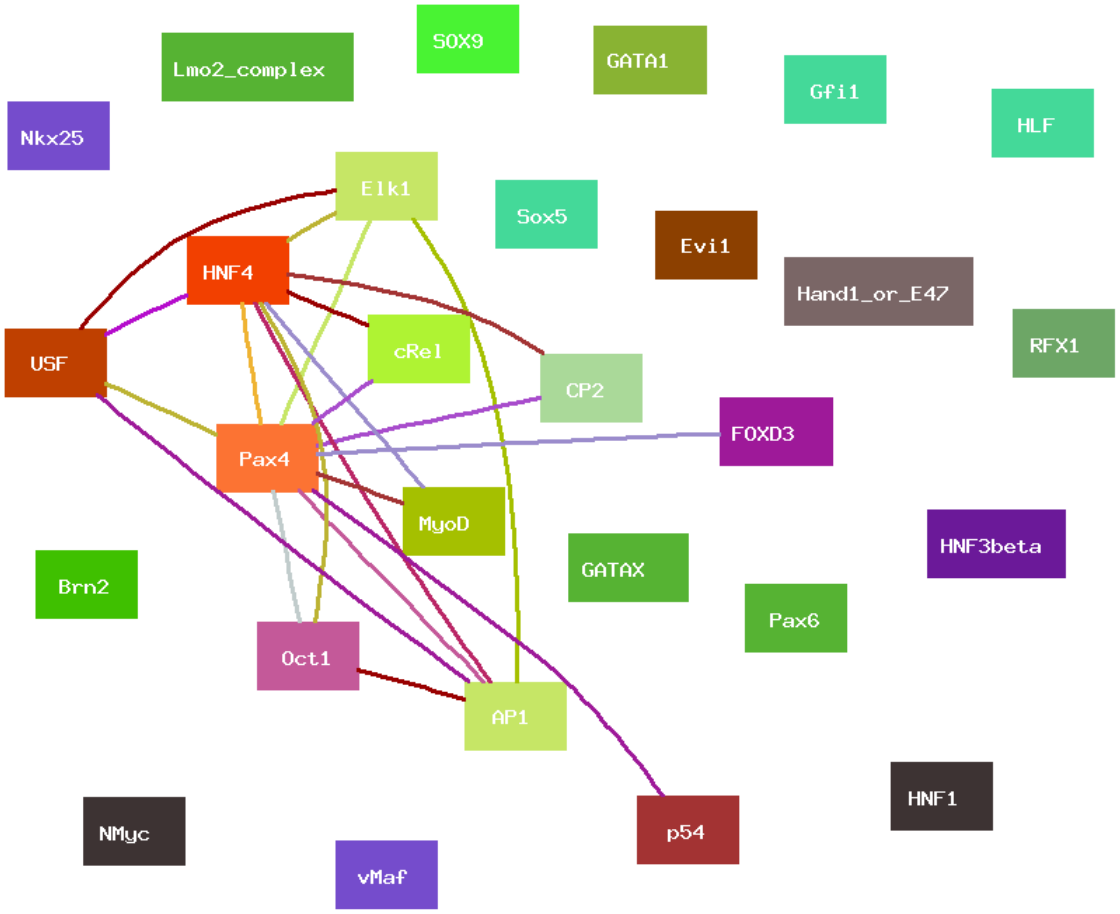


Figure 4.9: heart transcription factor co-occurrence network. The MATCH^{TM} output for the sequences obtained from the TCAT database visualized as transcription-factor co-occurrence network with color coding as mentioned on page 140, footnote 4.

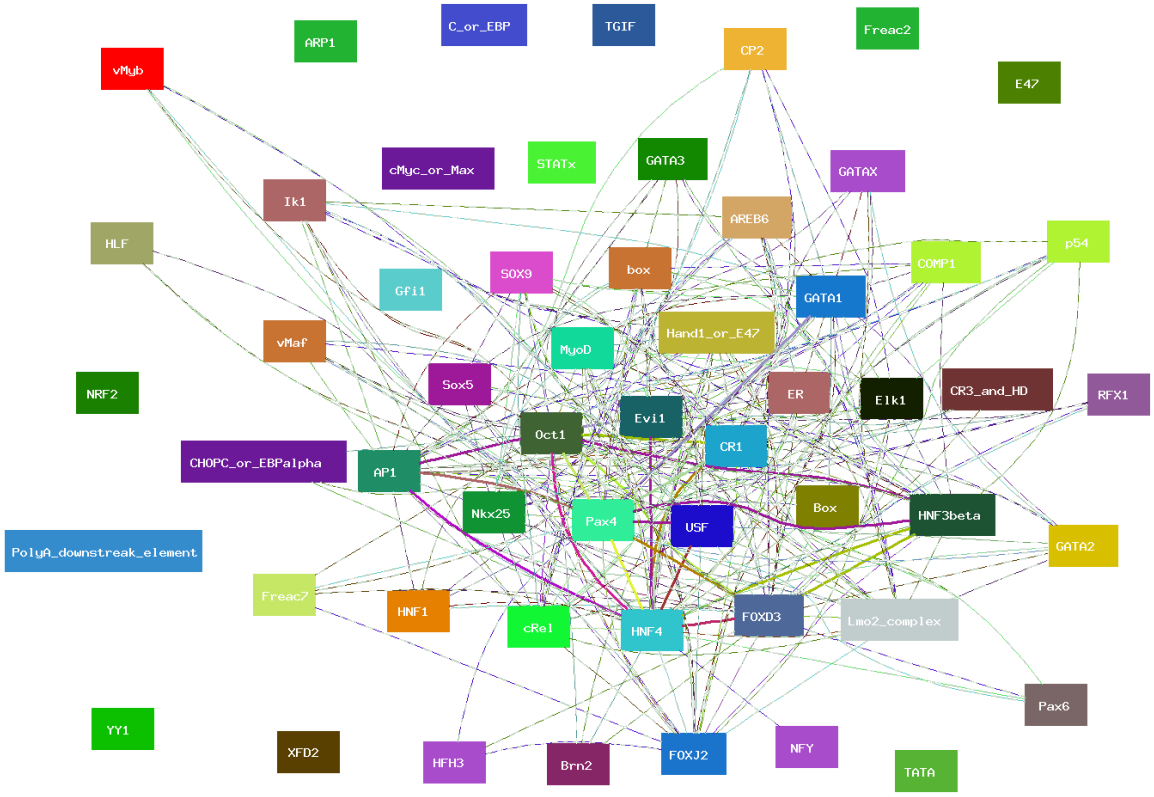


Figure 4.10: liver transcription factor co-occurrence network. The MATCH^{TM} output for the sequences obtained from the TCAT database visualized as transcription-factor co-occurrence network with color coding as mentioned on page 140, footnote 4.

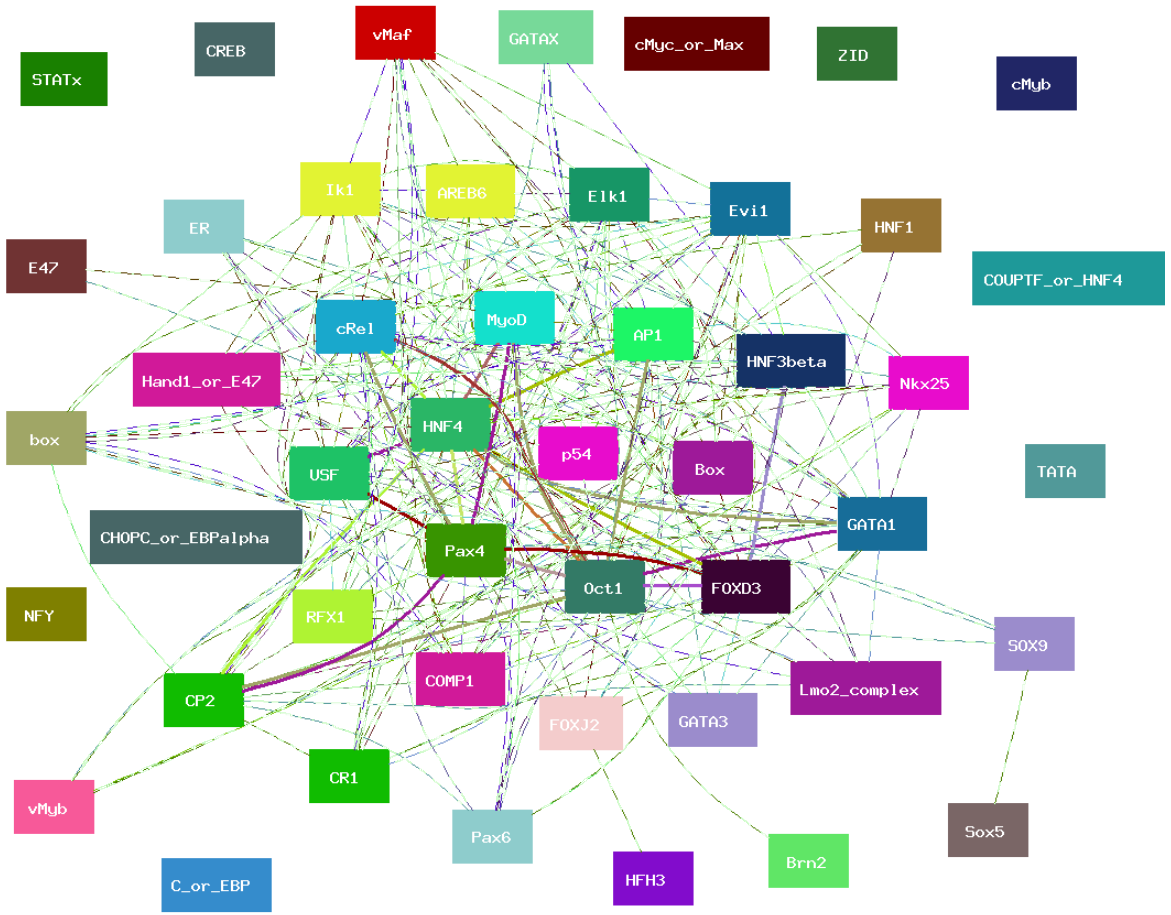


Figure 4.11: muscle transcription factor co-occurrence network The MATCH^{TM} output for the sequences obtained from the TCAT database visualized as transcription-factor co-occurrence network with color coding as mentioned on page 140, footnote 4.

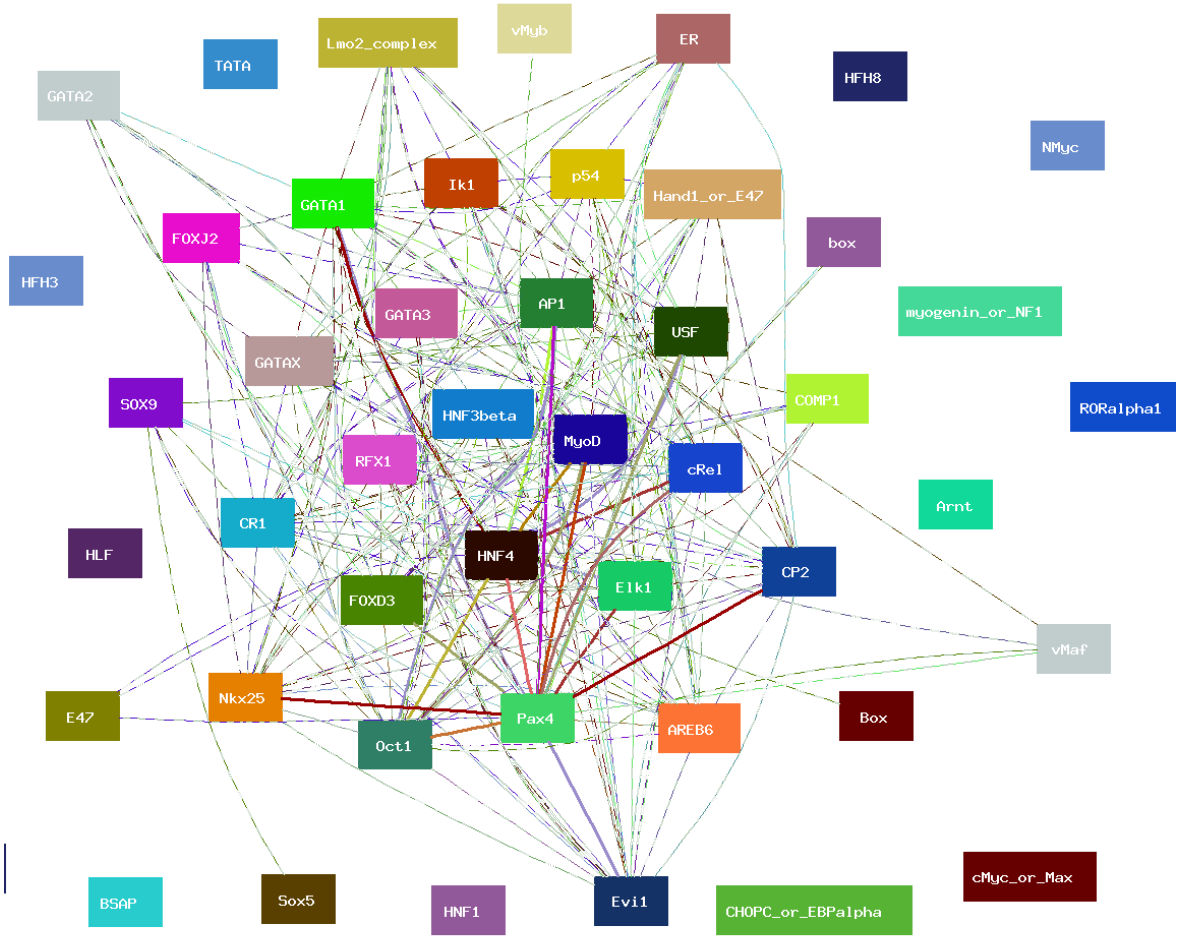


Figure 4.12: pancreas transcription factor co-occurrence network The MATCH^{TM} output for the sequences obtained from the TCAT database visualized as transcription-factor co-occurrence network with color coding as mentioned on page 140, footnote 4.

In Figures 4.8 through 4.12, it can be seen that the frequencies of co-occurrence of the known transcription factor binding sites in the upstream regions of the genes differ quite significantly across the tissues. Interestingly, though we see binding sites for nearly same transcription factors across all the promoters, that cannot be said to be true for the 'pairs' of transcription factor binding sites occurring together. In the above figures (4.8 through 4.12), each node represents a TFBS. Each link connecting two nodes means those to TFBSs occur together. The color of the node ranges from black to red. The closer the color of the node to red, more the occurrences of the TFBS in the given data set. Similarly closer the color of the link joining two nodes to red, higher the probability of their occurrence in the given data-set (for explanation of how the colors are derived see footnote 4 on page 140).

We therefore conclude that there is some information regarding the regulation of gene expression in the combination of potential TFBS on a promoter. Furthermore, it is also possible that some of this information is actually required or utilized in controlling the regulation of gene expression in tissue-specific manner. These networks are probably an indication of the complex regulatory machinery present in the cell(s) that takes care of the gene expression, which is very finely tuned in terms of its function.

4.4.4 Gene Networks

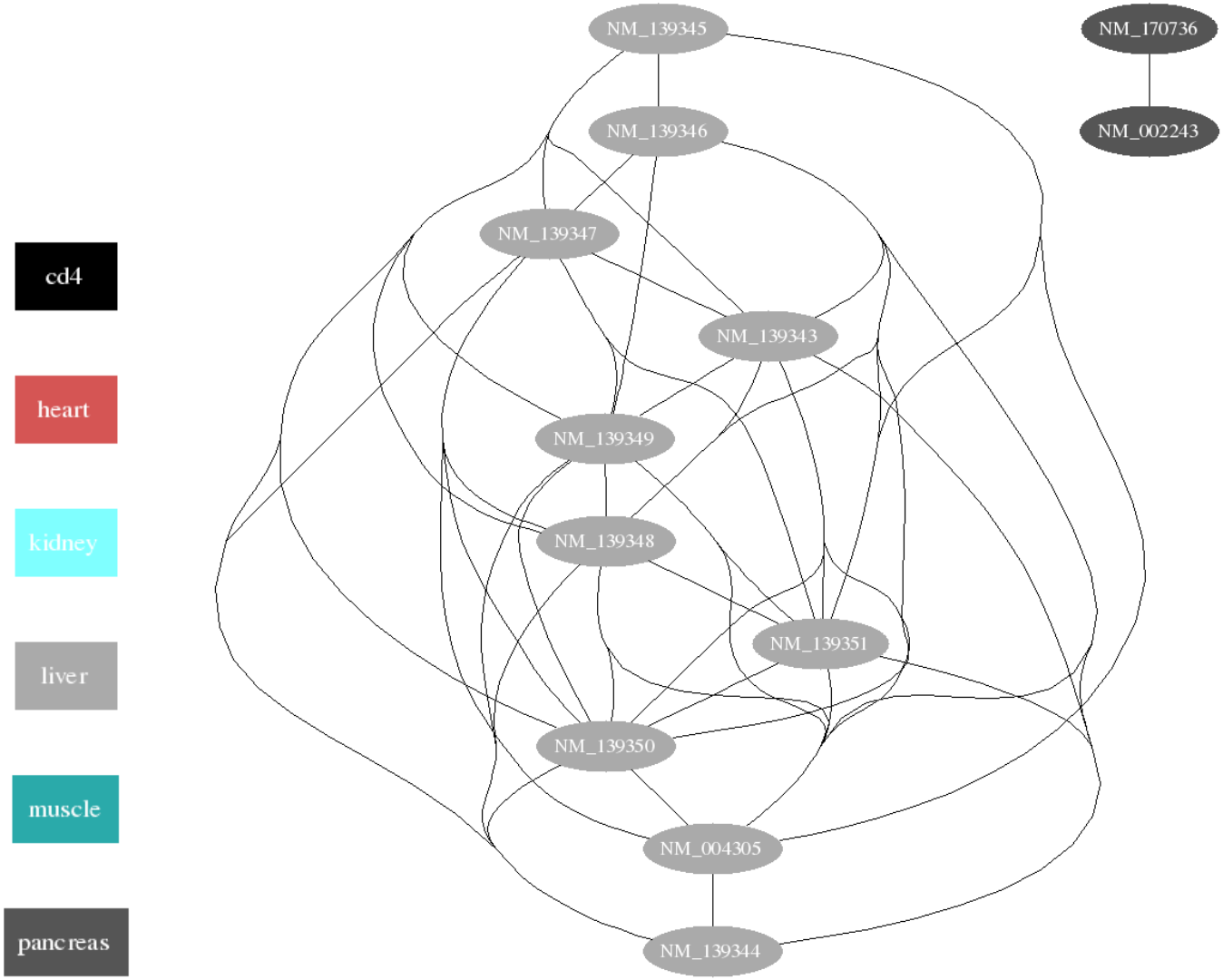
In this exercise, we used the same data as discussed earlier. Only we visualized the data on different criteria. This time we calculated number of common TFBS between each pair of genes. It was seen that the maximum number of transcription factors shared by any two genes was 39. In these figures, each node is name (Gene Accession Number of the Gene Database) of a gene. The use of gene symbol was avoided to reduce ambiguity in the visualization. The color of the node was fixed according to the prior information about the transcriptional status of the gene in tissues. The color-coding of the nodes is as mentioned in Figure 4.13(a). In figures 4.13 through 4.15, the networks are plotted according to the number of TFBS shared by a pair of promoter sequences. The number in the caption denotes the exact number of TFBS shared by two nodes in the graph for a connection between them to be plotted. The distribution of the number of connections is nearly *normal*. But most important information about tissue-specificity seems to be present in the tail region of the

distribution. This connectivity distribution is shown in Figure 4.16(a). It can be seen the number of TFBS shared between a pair can range from 0 to 39. When these plots are made using thresholds of 25 and above, an interesting picture emerges.

(a) Legend

(b) Gene network when the genes are plotted if and only if there are exactly 29 common transcription factors between a pair of genes.

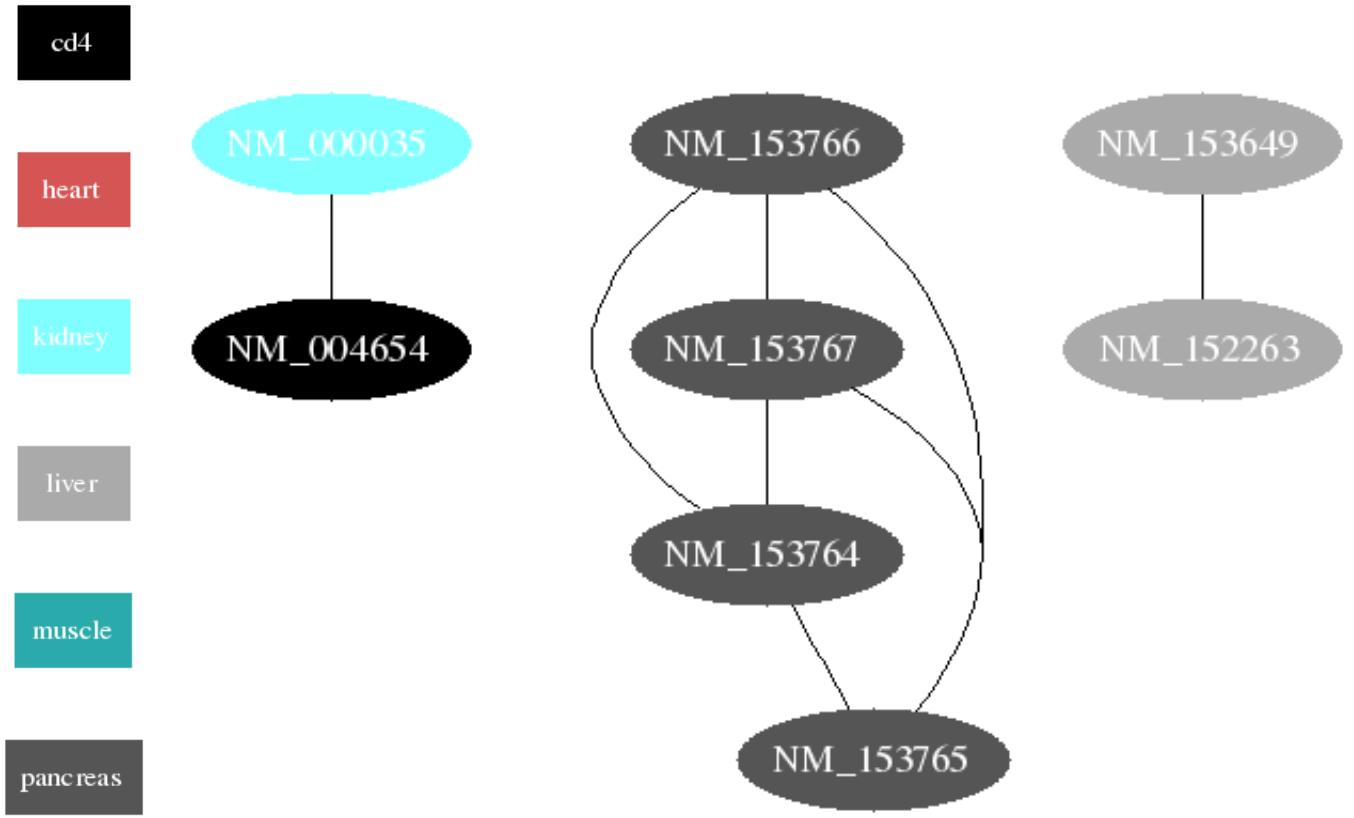
Figure 4.13: The panel a, shows the legend, the color coding used to denote genes expressed 'specifically' in a given tissue. b, shows the network of genes when all the genes are taken together for analysis and only those genes are plotted which have exactly 29 transcription factors in common.



(a) Legend

(b) Gene network 31 connections.

Figure 4.14: Gene network when the genes are plotted if and only if there are exactly 31 common transcription factors between a pair of genes (as acquired from the `TCAT` database (13)).



(a) Legend

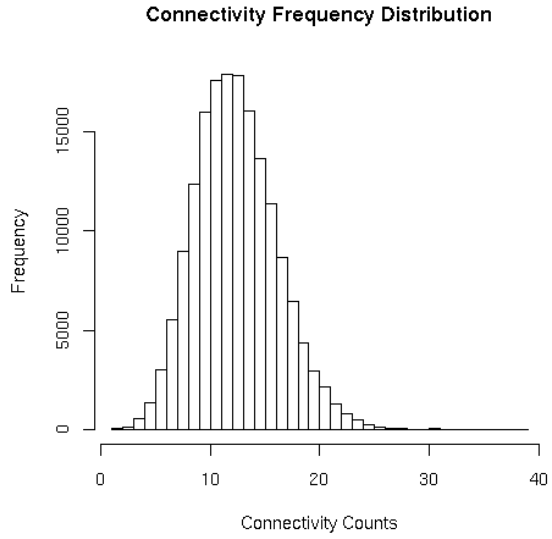
(b) Gene network 32 connections.

Figure 4.15: Gene network when the genes are plotted if and only if there are exactly 32 common transcription factors between a pair of genes (as acquired from the TCAT database (13)).

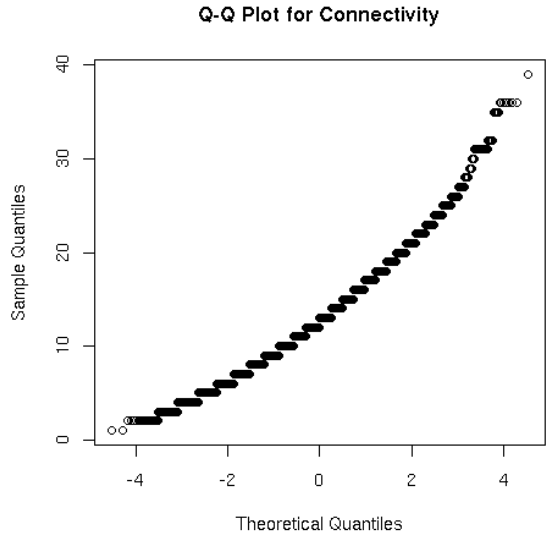
It can be seen in Figure 4.13, where the graph shows the connections if and only if each gene shares exactly 29 TFBS with another gene (in their upstream regions). Most connections are for genes which are either expressed specifically in liver or in kidney. It is possible that these genes are actually involved similar metabolic processes. When we actually performed the analysis to assign these genes to specific pathways, it was discovered that most of these promoters represent alternative forms (isoforms) of the same gene. Moreover, in some cases, it was also seen that the promoter is specific to a alternatively spliced product of the gene(s). This phenomenon has been described previously (23, 24). Further there is an isolated group of two genes, which are connected to each other and are not connected to any other genes, and both these are specific to the pancreas. This picture becomes clearer as the threshold for plotting is increased progressively. As shown in Figure 4.16(a), the distribution of the connectivity follows nearly *Normal Distribution*. To confirm this we plotted a normal q-q plot (Quantile-Quantile Plot) for the connectivity counts. The results are shown in Figure 4.16(b).

For a perfect normal distribution this graph will coincide with the diagonal, i.e., the theoretical quantiles are equal to the observed (sample) quantiles. However, as can be seen from Figure 4.16(a), the connectivity histogram is slightly skewed towards the right hand side. Furthermore, from the Figures 4.13(a) through 4.15, it is quite apparent that as the threshold connection for plotting the graphs is increased, the patterns that show up become more and more tissue/function specific. We can thus say that the information in the connectivity is in the right-hand tail of the graph.

This analysis however, is only a very basic analysis. It should be noted, that at least in the current analysis all the positional information about the TFBS has been totally ignored. There is lot of literature now available that demonstrates that the positional information in the TFBS distribution holds a lot of information concerning transcriptional regulation. However, considering all this information simultaneously for analysis can make the exercise quite unwieldy. Better methodologies will have to be developed (in addition to those described in the literature) to be able to deal with such complicated data in a elegant manner so that maximum information and hence knowledge can be obtained about the role of cis-regulatory regions in the control and regulation of transcription of genes.



(a) Distribution of shared TFBS between each pair of genes in the data set.



(b) The connectivity distribution follows nearly Normal Distribution.

Figure 4.16: Distribution of the number of shared transcription factor binding sites (TFBS) across the dataset is nearly *normal*.

4.5 Discussion, Conclusions and Future Perspectives

From the results presented in the preceding pages following interesting and important observations can be made:

- Similar transcription factors have binding sites on various promoters, implying that a very small set of transcription factors can regulate a large number of genes.
- It is the combinations of the TFBS on the promoters that gives the promoter a distinctive identity in terms of its expression. It is reasonable to believe that occurrence of a set of TFBS on a set of promoters will probably lead to their co-expression (25, 26). Though there are reports demonstrating that occurrence of a TFBS, or even its binding of a TF to its cognate site does not necessarily mean that the TF affects the expression of the gene (27, 28).
- The co-occurrence of TFBS along the promoter does vary considerably (at least visually) as a function of tissue. However additional statistical analysis of the graphs would be necessary to highlight the systematic differences in terms of TFBS.
- It should be noted that in all the analyses with TFBS the positional information has not been included. The position of the TFBSs with respect to the TSS and with respect to each other does seem to play important role in the regulation of downstream genes (29). However, addition of this aspect into the analyses of TFBS data increases the complexity of the analyses immensely and few more statistical analytical techniques will have to be developed to account for such complexity. Future studies will have to take this into consideration.

The major bottleneck in such analyses is the availability of proper input sequences or curated promoter databases. Not many genes have been characterized in terms of their tissue-specic expression. Also, the TFBS analysis was carried out using the `MATCH` program, that contains only few TFBS weight matrices. In principle it should be possible to curate the microarray data submitted to the `GEO` that is now available in the public domain to obtain tissue-specically expressed/repressed genes. This information can be utilized to obtain their respective promoter sequences from other well-curated databases such as the `Genomatix Promoter Database`. The analyses on such sequences would lead to better insights in to

understanding of the gene-regulatory networks and other cis-regulatory modules that play important role in regulated and coordinated expression of genes.

Additionally, from the distribution of the co-occurring TFBS in the promoters (without their relative or absolute positional information) it seems that presence/absence/co-occurrence information alone may provide a basic level of understanding about general regulatory principles for gene expression. The presented results also highlight the complexity of problem of deciphering regulatory networks from primary genomic sequence information alone, and at the same time hint that it may, in principle, be possible to address this problem using special statistical techniques.

References

[1] N. I. Gershenzon, E. N. Trifonov, and I. P. Ioshikhes. The features of Drosophila core promoters revealed by statistical analysis. *BMC Genomics*, 7:161, 2006.

[2] A. Stein and M. Bina. A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment. *Nucleic Acids Res*, 27(3):848–53, 1999.

[3] V. Beisvag, F. K. Junge, H. Bergum, L. Jolsum, S. Lydersen, C. C. Gunther, H. Ramampiaro, M. Langaas, A. K. Sandvik, and A. Laegreid. GeneTools–application for functional annotation and statistical hypothesis testing. *BMC Bioinformatics*, 7:470, 2006.

[4] N. Harada, T. Utsumi, and Y. Takagi. Tissue-specific expression of the human aromatase cytochrome P-450 gene by alternative use of multiple exons 1 and promoters, and switching of tissue-specific exons 1 in carcinogenesis. *Proc Natl Acad Sci U S A*, 90(23):11312–6, 1993.

[5] J. Adjaye. Whole-genome approaches for large-scale gene identification and expression analysis in mammalian preimplantation embryos. *Reprod Fertil Dev*, 17(1-2):37–45, 2005.

[6] M. Shklar, L. Strichman-Almashanu, O. Shmueli, M. Shmoish, M. Safran, and D. Lancet. GeneTide–Terra Incognita Discovery Endeavor: a new transcriptome focused member of the GeneCards/GeneNote suite of databases. *Nucleic Acids Res*, 33(Database issue):D556–61, 2005.

[7] O. Shmueli, S. Horn-Saban, V. Chalifa-Caspi, M. Shmoish, R. Ophir, H. Benjamin-Rodrig, M. Safran, E. Domany, and D. Lancet. GeneNote: whole genome expression profiles in normal human tissues. *C R Biol*, 326(10-11):1067–72, 2003.

[8] Y. Li, K. K. Lee, S. Walsh, C. Smith, S. Hadingham, K. Sorefan, G. Cawley, and M. W. Bevan. Establishing glucose- and ABA-regulated transcription networks in Arabidopsis by microarray analysis and promoter classification using a Relevance Vector Machine. *Genome Res*, 16(3):414–27, 2006.

[9] K. Frech, K. Quandt, and T. Werner. Muscle actin genes: a first step towards computational classification of tissue specific promoters. *In Silico Biol*, 1(1):29–38, 1998.

[10] I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, and O. Shmueli. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*, 21(5):650–9, 2005.

[11] J. Schug, W. P. Schuller, C. Kappen, J. M. Salbaum, M. Bucan, and C. J. Stoeckert, Jr. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol*, 6(4):R33, 2005.

[12] P. C. Fitzgerald, A. Shlyakhtenko, A. A. Mir, and C. Vinson. Clustering of dna sequences in human promoters. *Genome Res*, 14:1565–74, 2004.

[13] A. D. Smith, P. Sumazin, Z. Xuan, and M. Q. Zhang. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. *Proc Natl Acad Sci U S A*, 103(16):6275–80, 2006.

[14] C. Zhang, Z. Xuan, S. Otto, J. R. Hover, S. R. McCorkle, G. Mandel, and M. Q. Zhang. A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. *Nucleic Acids Res*, 34(8):2238–46, 2006.

[15] X. Xie, T. S. Mikkelsen, A. Gnirke, K. Lindblad-Toh, M. Kellis, and E. S. Lander. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci U S A*, 104(17):7145–50, 2007.

[16] D. Xie, J. Cai, N. Y. Chia, H. H. Ng, and S. Zhong. Cross-species de novo identification of cis-regulatory modules with GibbsModule: Application to gene regulation in embryonic stem cells. *Genome Res*, 18(8):1325–35, 2008.

[17] R. Knuppel, P. Dietze, W. Lehnberg, K. Frech, and E. Wingender. TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. *J Comput Biol*, 1(3):191–8, 1994.

[18] G. B. Fogel, D. G. Weekes, G. Varga, E. R. Dow, A. M. Craven, H. B. Harlow, E. W. Su, J. E. Onyia, and C. Su. A statistical analysis of the TRANSFAC database. *Biosystems*, 81(2):137–54, 2005.

[19] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–44, 2005.

[20] X. Li, S. Rao, W. Jiang, C. Li, Y. Xiao, Z. Guo, Q. Zhang, L. Wang, L. Du, J. Li, L. Li, T. Zhang, and Q. K. Wang. Discovery of Time-Delayed Gene Regulatory Networks based on temporal gene expression profiling. *BMC Bioinformatics*, 7:26, 2006.

[21] D.B. Searls. The language of genes. *Nature*, 420(6912):211–217, 2002.

[22] Kel A. E., Gossling E, Reuter I., Cheremushkin E., Kel-Margulis O. V., and Wingender E. Match: A tool for searching transcription factor binding sites in dna sequences. *Nucleic Acids Res.*, 31(13):3576–9, 2003.

[23] A. Kpebe and L. Rabinow. Alternative promoter usage generates multiple evolutionarily conserved isoforms of Drosophila DOA kinase. *Genesis*, 46(3):132–43, 2008.

[24] P. D. Ellis, C. W. Smith, and P. Kemp. Regulated tissue-specific alternative splicing of enhanced green fluorescent protein transgenes conferred by alpha-tropomyosin regulatory elements in transgenic mice. *J Biol Chem*, 279(35):36660–9, 2004.

[25] J. L. Gomez-Skarmeta, I. Rodriguez, C. Martinez, J. Culi, D. Ferres-Marco, D. Beamonte, and J. Modolell. Cis-regulation of achaete and scute: shared enhancer-like elements drive their coexpression in proneural clusters of the imaginal discs. *Genes Dev*, 9(15):1869–82, 1995.

[26] V. J. Makeev, A. P. Lifanov, A. G. Nazina, and D. A. Papatsenko. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res*, 31(20):6016–26, 2003.

[27] D. A. Papatsenko, V. J. Makeev, A. P. Lifanov, M. Regnier, A. G. Nazina, and C. Desplan. Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers. *Genome Res*, 12(3):470–81, 2002.

[28] M. L. Allende, M. Manzanares, J. J. Tena, C. G. Feijoo, and J. L. Gomez-Skarmeta. Cracking the genome's second code: enhancer detection by combined phylogenetic footprinting and transgenic fish and frog embryos. *Methods*, 39(3):212–9, 2006.

[29] J. Lu, L. Luo, and Y. Zhang. Distance conservation of transcription regulatory motifs in human promoters. *Comput Biol Chem*, 2008.

Appendix A

Motif Models (PSPM)s

A.1 Background Model MD0

Position	A	C	G	T	Motif
1	0.000000	1.000000	0.000000	0.000000	C
2	0.000000	1.000000	0.000000	0.000000	C
3	0.000000	0.002381	0.000000	0.997619	T
4	0.000000	1.000000	0.000000	0.000000	C
5	0.626143	0.075790	0.191495	0.106572	A
6	0.000000	0.005125	0.994875	0.000000	G
7	0.000000	1.000000	0.000000	0.000000	C
8	0.000000	1.000000	0.000000	0.000000	C
9	0.000000	0.004610	0.000000	0.995390	T
10	0.000000	1.000000	0.000000	0.000000	C
11	0.000000	1.000000	0.000000	0.000000	C
12	0.000000	0.810820	0.000000	0.189180	C

Position	A	C	G	T	Motif
1	0.538217	0.000000	0.158070	0.303713	A
2	0.000000	0.000000	1.000000	0.000000	G
3	0.000000	0.983162	0.016838	0.000000	C
4	0.010910	0.156299	0.000000	0.832791	T
5	0.000000	0.000000	0.995338	0.004662	G
6	0.004165	0.000000	0.995835	0.000000	G
7	0.000000	0.000000	1.000000	0.000000	G
8	0.661265	0.260130	0.056193	0.022412	A
9	0.092311	0.205035	0.221105	0.481549	T
10	0.005485	0.091466	0.010972	0.892077	T
11	0.672997	0.000000	0.327003	0.000000	A
12	0.000000	0.667271	0.332729	0.000000	C
13	0.683245	0.000000	0.012630	0.304125	A
14	0.000000	0.000000	1.000000	0.000000	G
15	0.000000	0.000000	1.000000	0.000000	G
16	0.000000	0.947974	0.039384	0.012641	C

Position	A	C	G	T	Motif
1	0.000000	0.835510	0.164490	0.000000	C
2	0.148725	0.210882	0.000000	0.640393	T
3	0.000000	0.988865	0.003705	0.007431	C
4	0.000000	1.000000	0.000000	0.000000	C
5	0.761231	0.022866	0.000000	0.215904	G
6	0.000000	0.007707	0.978594	0.013699	G
7	0.000000	1.000000	0.000000	0.000000	C
8	0.000000	1.000000	0.000000	0.000000	C
9	0.000697	0.073898	0.000000	0.925405	T
10	0.000000	0.225980	0.735969	0.038051	G
11	0.154508	0.086047	0.743004	0.016440	G
12	0.031936	0.222319	0.716750	0.028995	G

Position	A	C	G	T	Motif
1	0.991486	0.000008	0.008506	0.000000	A
2	0.585790	0.002876	0.411334	0.000000	A
3	0.979513	0.000000	0.020487	0.000000	A
4	0.007875	0.970595	0.005147	0.016383	C
5	0.007220	0.585244	0.001927	0.405608	C
6	0.000000	0.949862	0.008877	0.041261	C
7	0.007349	0.662900	0.003112	0.326638	C
8	0.350546	0.003319	0.642873	0.003262	G
9	0.004144	0.009398	0.000000	0.986459	T
10	0.003758	0.968884	0.008529	0.018829	C
11	0.001355	0.010481	0.000986	0.987178	T
12	0.003505	0.956097	0.016285	0.024113	C
13	0.401711	0.017393	0.000000	0.580896	T
14	0.980803	0.000000	0.019197	0.000000	A
15	0.413043	0.550556	0.011515	0.024886	C
16	0.471517	0.022896	0.000000	0.505587	T
17	0.949882	0.011610	0.024462	0.014046	A
18	0.974961	0.013909	0.001973	0.009158	A
19	0.970540	0.011631	0.000000	0.017830	A
20	0.955723	0.013573	0.014061	0.016642	A
21	0.961790	0.003524	0.023388	0.011298	A

Position	A	C	G	T	Motif
1	0.000000	0.781620	0.000000	0.218380	C
2	0.000000	1.000000	0.000000	0.000000	C
3	0.000000	0.885937	0.000000	0.114063	C
4	0.000000	1.000000	0.000000	0.000000	C
5	0.297057	0.000000	0.000000	0.702943	T
6	0.000000	0.628568	0.295797	0.075635	C
7	0.000000	1.000000	0.000000	0.000000	C
8	0.000000	0.905960	0.000000	0.094040	C
9	0.000000	0.854798	0.000000	0.145202	C

A.2 Background Model RP0

Position	A	C	G	T	Motif
1	0.956760	0.000000	0.038748	0.004491	A
2	0.020433	0.010910	0.965295	0.003361	G
3	0.002358	0.009198	0.982174	0.006270	G
4	0.000000	0.926988	0.020251	0.052761	C
5	0.101409	0.249507	0.122250	0.526834	T
6	0.052659	0.009746	0.937594	0.000000	G
7	0.955717	0.002606	0.041677	0.000000	A
8	0.002575	0.000000	0.997425	0.000000	G
9	0.003826	0.012044	0.984130	0.000000	G
10	0.017085	0.717311	0.013581	0.252024	C
11	0.504606	0.000000	0.290189	0.205204	A
12	0.016873	0.000000	0.977071	0.006056	G
13	0.012017	0.195763	0.792220	0.000000	G
14	0.692381	0.146476	0.042748	0.118395	A
15	0.050850	0.005019	0.944132	0.000000	G
16	0.391327	0.020197	0.393858	0.194618	A

Position	A	C	G	T	Motif
1	0.134576	0.662374	0.060458	0.142592	C
2	0.123218	0.122109	0.744481	0.010192	G
3	0.000000	0.997638	0.000000	0.002362	C
4	0.000000	1.000000	0.000000	0.000000	C
5	0.329825	0.048401	0.000000	0.621774	T
6	0.000000	0.416692	0.583308	0.000000	G
7	0.004225	0.399189	0.000000	0.596586	T
8	0.777447	0.058230	0.152383	0.011939	A
9	0.293105	0.346473	0.308265	0.052157	C
10	0.022284	0.143941	0.282746	0.551028	T
11	0.000000	1.000000	0.000000	0.000000	C
12	0.000000	0.992444	0.000000	0.007556	C
13	0.014764	0.969957	0.000000	0.015279	C
14	0.732431	0.026301	0.196716	0.044552	A
15	0.000000	0.055422	0.944578	0.000000	G
16	0.000000	0.975360	0.019970	0.004670	C

Position	A	C	G	T	Motif
1	0.000000	0.979883	0.020117	0.000000	C
2	0.000000	1.000000	0.000000	0.000000	C
3	0.610808	0.097526	0.000000	0.291666	A
4	0.000000	0.096589	0.895898	0.007513	G
5	0.000000	1.000000	0.000000	0.000000	C
6	0.000000	1.000000	0.000000	0.000000	C
7	0.034341	0.115176	0.000000	0.850483	T
8	0.000000	0.215248	0.694328	0.090424	G
9	0.124171	0.096795	0.734056	0.044978	G
10	0.000000	0.374974	0.616829	0.008196	G
11	0.000000	0.996599	0.000000	0.003401	C
12	0.424087	0.475768	0.084425	0.015720	C

Position	A	C	G	T	Motif
1	0.083963	0.473281	0.275862	0.166894	C
2	0.143632	0.563570	0.107389	0.185409	C
3	0.007011	0.802949	0.043452	0.146589	C
4	0.039731	0.834651	0.000000	0.125618	C
5	0.036337	0.850051	0.000000	0.113612	C
6	0.000000	1.000000	0.000000	0.000000	C
7	0.619365	0.000000	0.005350	0.375285	A
8	0.000000	0.282603	0.689035	0.028362	G
9	0.000000	0.859319	0.133544	0.007137	C
10	0.006614	0.903846	0.000000	0.089540	C
11	0.019770	0.684773	0.000000	0.295457	C
12	0.038075	0.816854	0.042751	0.102320	C

Position	A	C	G	T	Motif
1	0.043108	0.209453	0.714437	0.033003	G
2	0.007746	0.953029	0.004519	0.034707	C
3	0.000000	1.000000	0.000000	0.000000	C
4	0.607617	0.047270	0.000000	0.345114	A
5	0.000000	0.643910	0.346607	0.009482	C
6	0.046411	0.561404	0.032680	0.359504	C
7	0.499418	0.131947	0.319664	0.048970	A
8	0.199636	0.340867	0.155341	0.304156	C
9	0.168306	0.213835	0.394292	0.223567	G
10	0.002949	0.941474	0.032560	0.023017	C
11	0.000000	0.935060	0.000000	0.064940	C
12	0.050143	0.684494	0.000000	0.265362	C
13	0.398819	0.028978	0.456463	0.115740	G
14	0.004603	0.089047	0.896258	0.010092	G
15	0.000000	0.969485	0.015873	0.014642	C
16	0.075852	0.563219	0.000000	0.360929	C

A.3 Background Model RP1

Position	A	C	G	T	Motif
1	0.156978	0.793977	0.043426	0.005619	C
2	0.008671	0.929793	0.029796	0.031740	C
3	0.007117	0.085103	0.051707	0.856073	T
4	0.001740	0.017467	0.980793	0.000000	G
5	0.000016	0.254045	0.000000	0.745939	T
6	0.859206	0.017192	0.123602	0.000000	A
7	0.538743	0.045410	0.415847	0.000000	A
8	0.000000	0.154498	0.089721	0.755781	T
9	0.000000	0.970673	0.029327	0.000000	C
10	0.013241	0.789515	0.031577	0.165667	C
11	0.000000	0.931472	0.058304	0.010223	C
12	0.708330	0.238420	0.049013	0.004237	A
13	0.000000	0.079339	0.920661	0.000000	G
14	0.016171	0.945671	0.028826	0.009332	C
15	0.110454	0.203732	0.062840	0.622974	T
16	0.602801	0.201507	0.032712	0.162980	A
17	0.000000	0.917636	0.000000	0.082364	C
18	0.015208	0.215879	0.038891	0.730022	T
19	0.000000	0.670937	0.025423	0.303640	C
20	0.102485	0.043202	0.854313	0.000000	G
21	0.000000	0.019824	0.974749	0.005428	G
22	0.018386	0.120467	0.861147	0.000000	G

Position	A	C	G	T	Motif
1	0.121219	0.019308	0.859473	0.000000	G
2	0.033761	0.029287	0.936940	0.000012	G
3	0.000000	0.964718	0.031182	0.004100	C
4	0.012939	0.750000	0.052159	0.184902	C
5	0.144935	0.004222	0.841169	0.009674	G
6	0.371074	0.041686	0.587240	0.000000	G
7	0.014401	0.017868	0.967731	0.000000	G
8	0.012778	0.464479	0.419555	0.103189	C
9	0.129921	0.352112	0.455445	0.062521	G
10	0.119538	0.363388	0.317130	0.199944	C
11	0.097120	0.009767	0.883347	0.009767	G
12	0.013325	0.046246	0.935134	0.005296	G
13	0.076840	0.201965	0.022165	0.699031	T
14	0.028132	0.015525	0.956344	0.000000	G
15	0.051184	0.034345	0.914471	0.000000	G
16	0.363152	0.578021	0.046561	0.012265	C
17	0.038867	0.031182	0.113924	0.816026	T
18	0.009767	0.905099	0.080210	0.004925	C
19	0.759824	0.033141	0.197684	0.009352	A
20	0.001499	0.832235	0.084996	0.081270	C
21	0.070998	0.156351	0.645599	0.127052	G
22	0.159219	0.540452	0.078060	0.222269	C

Position	A	C	G	T	Motif
1	0.245682	0.044876	0.709441	0.000000	G
2	0.006581	0.513269	0.480150	0.000000	C
3	0.039127	0.672293	0.266343	0.022237	C
4	0.025632	0.454533	0.363662	0.156173	C
5	0.181231	0.121741	0.605264	0.091763	G
6	0.059004	0.057949	0.870968	0.012079	G
7	0.081057	0.107133	0.811810	0.000000	G
8	0.299365	0.206370	0.403751	0.090515	G
9	0.044319	0.113206	0.713781	0.128695	G
10	0.054588	0.101078	0.811242	0.033093	G
11	0.019251	0.755647	0.210509	0.014593	C
12	0.395882	0.109358	0.372784	0.121977	A
13	0.000000	0.164978	0.835022	0.000000	G
14	0.334077	0.149541	0.375121	0.141260	A
15	0.000000	0.279508	0.706017	0.014475	G
16	0.017590	0.189427	0.792983	0.000000	G

Position	A	C	G	T	Motif
1	0.336059	0.003923	0.652629	0.007389	G
2	0.045653	0.003730	0.948580	0.002037	G
3	0.000000	0.035065	0.004383	0.960552	T
4	0.000000	0.446977	0.010529	0.542495	T
5	0.000000	0.020278	0.000000	0.979722	T
6	0.001936	0.722522	0.004653	0.270889	C
7	0.569753	0.038311	0.330946	0.060990	G
8	0.003306	0.946750	0.000000	0.049944	C
9	0.010406	0.510910	0.011433	0.467251	C
10	0.524648	0.309704	0.134524	0.031124	A
11	0.016941	0.024034	0.004399	0.954626	T
12	0.050942	0.018782	0.911436	0.018840	G
13	0.000000	0.007758	0.004428	0.987815	T
14	0.006979	0.245679	0.000000	0.747342	T
15	0.270492	0.018157	0.678078	0.033273	G
16	0.004827	0.469124	0.496049	0.030000	G
17	0.000000	0.847539	0.006337	0.146124	C
18	0.000000	0.959209	0.007556	0.033235	C
19	0.956304	0.000000	0.027916	0.015779	A
20	0.015397	0.004383	0.978182	0.002037	G
21	0.034164	0.022917	0.936710	0.006209	G
22	0.180232	0.787318	0.013022	0.019427	C
23	0.000000	0.040227	0.000000	0.959773	T
24	0.042764	0.008119	0.942187	0.006930	G
25	0.000013	0.000000	0.999974	0.000013	G
26	0.328967	0.007933	0.017695	0.645405	T
27	0.012581	0.638755	0.314497	0.034167	C
28	0.007790	0.009691	0.019227	0.963292	T

Position	A	C	G	T	Motif
1	0.000000	0.907979	0.000000	0.092021	C
2	0.001585	0.907406	0.057061	0.033948	C
3	0.000000	0.970637	0.006611	0.022752	C
4	0.987074	0.000000	0.012926	0.000000	A
5	0.931514	0.000000	0.068486	0.000000	A
6	0.885805	0.000000	0.056662	0.057533	A
7	0.046868	0.000000	0.870483	0.082648	G
8	0.122994	0.000000	0.000013	0.876993	T
9	0.005200	0.000000	0.994800	0.000000	G
10	0.000000	0.945169	0.000000	0.054831	C
11	0.000000	0.051722	0.000000	0.948278	T
12	0.033016	0.000000	0.966984	0.000000	G
13	0.091449	0.000000	0.908551	0.000000	G
14	0.006446	0.000000	0.993554	0.000000	G
15	1.000000	0.000000	0.000000	0.000000	A
16	0.000000	0.022535	0.000000	0.977465	T
17	0.000000	0.040432	0.000000	0.959568	T
18	0.968414	0.000000	0.031586	0.000000	A
19	0.000000	0.954135	0.000000	0.045865	C
20	0.959353	0.000000	0.040647	0.000000	A

A.4 Background Model RP2

Position	A	C	G	T	Motif
1	0.572993	0.017593	0.385036	0.024377	A
2	0.000000	0.028938	0.971062	0.000000	G
3	0.000000	0.995838	0.000000	0.004162	C
4	0.000000	0.555045	0.034125	0.410830	C
5	0.004018	0.000000	0.995982	0.000000	G
6	0.090452	0.020575	0.888973	0.000000	G
7	0.000000	0.009914	0.990086	0.000000	G
8	0.376954	0.493545	0.107307	0.022194	C
9	0.070960	0.350834	0.458898	0.119308	G
10	0.000000	0.260462	0.152208	0.587329	T
11	0.360225	0.000000	0.639775	0.000000	G
12	0.000000	0.353070	0.646930	0.000000	G
13	0.351323	0.122254	0.029918	0.496506	T
14	0.000000	0.013317	0.986683	0.000000	G
15	0.009914	0.025614	0.964473	0.000000	G
16	0.084737	0.854905	0.020587	0.039772	C

Position	A	C	G	T	Motif
1	0.308050	0.040959	0.650991	0.000000	G
2	0.000000	0.978039	0.018982	0.002979	C
3	0.000000	0.997112	0.000000	0.002888	C
4	0.000000	0.024924	0.082608	0.892468	T
5	0.000000	1.000000	0.000000	0.000000	C
6	0.339560	0.244659	0.415781	0.000000	G
7	0.030291	0.009660	0.951475	0.008573	G
8	0.000000	0.973254	0.007057	0.019690	C
9	0.000000	0.980680	0.019320	0.000000	C
10	0.009660	0.060427	0.023866	0.906047	T
11	0.000000	0.998141	0.001859	0.000000	C
12	0.000000	0.963864	0.036136	0.000000	C

Position	A	C	G	T	Motif
1	0.022966	0.875024	0.006539	0.095471	C
2	0.858856	0.005398	0.116924	0.018822	A
3	0.022699	0.013569	0.957822	0.005910	G
4	0.097844	0.022961	0.003161	0.876033	T
5	0.027193	0.013184	0.856609	0.103013	G
6	0.884742	0.097844	0.015841	0.001574	A
7	0.096016	0.012703	0.886716	0.004565	G
8	0.005169	0.971876	0.007280	0.015675	C
9	0.048866	0.662537	0.034937	0.253659	C
10	0.190327	0.032107	0.693363	0.084203	G
11	0.830593	0.016044	0.110137	0.043227	A
12	0.106215	0.000000	0.889497	0.004288	G
13	0.867845	0.006829	0.113747	0.011579	A
14	0.000000	0.049069	0.080547	0.870384	T
15	0.034129	0.584028	0.058499	0.323344	G
16	0.248256	0.142873	0.604223	0.004649	G
17	0.091362	0.491902	0.062734	0.354002	C
18	0.231468	0.000000	0.749474	0.019058	G
19	0.022717	0.862811	0.098698	0.015773	C
20	0.097844	0.886738	0.005512	0.009905	C
21	0.896594	0.000000	0.103406	0.000000	A
22	0.005715	0.730844	0.022141	0.241301	C
23	0.004459	0.006288	0.016240	0.973013	T
24	0.019077	0.103420	0.864403	0.013100	G
25	0.026160	0.842432	0.090363	0.041046	C
26	0.938148	0.025311	0.036541	0.000000	A
27	0.013924	0.831515	0.119369	0.035192	C

Position	A	C	G	T	Motif
1	0.000000	0.985879	0.000000	0.014121	C
2	0.000000	0.999989	0.000000	0.000011	C
3	0.000000	0.006759	0.000000	0.993241	T
4	0.013931	0.000000	0.986069	0.000000	G
5	0.000000	0.000000	0.000000	1.000000	T
6	0.998463	0.001537	0.000000	0.000000	A
7	0.906616	0.000000	0.093384	0.000000	A
8	0.000000	0.000000	0.000000	1.000000	T
9	0.000000	0.988383	0.000000	0.011617	C
10	0.000000	0.982760	0.000017	0.017223	C
11	0.000000	0.976962	0.000000	0.023038	C
12	1.000000	0.000000	0.000000	0.000000	A
13	0.049510	0.000000	0.950490	0.000000	G
14	0.010073	0.987024	0.000000	0.002903	C

Position	A	C	G	T	Motif
1	0.011467	0.033108	0.069412	0.886014	T
2	0.000000	0.432088	0.049632	0.518280	T
3	0.000000	0.916697	0.045097	0.038207	C
4	0.108322	0.053825	0.135185	0.702667	T
5	0.000000	0.943416	0.045412	0.011173	C
6	0.000000	0.924988	0.075012	0.000000	C
7	0.000000	0.313816	0.041480	0.644704	T
8	0.177608	0.030241	0.769560	0.022592	G
9	0.000000	0.974892	0.007349	0.017758	C
10	0.007446	0.948263	0.023583	0.020707	C
11	0.000016	0.094381	0.016651	0.888951	T
12	0.010061	0.782371	0.042751	0.164816	C
13	0.693130	0.052868	0.240901	0.013101	A
14	0.020995	0.045018	0.928577	0.005410	G
15	0.016200	0.929810	0.016651	0.037340	C
16	0.006826	0.945990	0.005410	0.041774	C
17	0.018332	0.052446	0.036772	0.892450	T
18	0.000000	0.966075	0.032079	0.001846	C
19	0.001544	0.951735	0.040664	0.006057	C
20	0.010061	0.730761	0.034493	0.224685	C
21	0.399777	0.093309	0.506914	0.000000	G

A.5 Background Model RP3

Position	A	C	G	T	Motif
1	0.287091	0.024854	0.688055	0.000000	G
2	0.000000	0.987043	0.009904	0.003054	C
3	0.000000	0.997039	0.000000	0.002961	C
4	0.078033	0.013127	0.064792	0.844048	T
5	0.000000	1.000000	0.000000	0.000000	C
6	0.251898	0.296326	0.451776	0.000000	G
7	0.039855	0.009904	0.941452	0.008789	G
8	0.000000	0.981472	0.000000	0.018528	C
9	0.000000	0.894133	0.105867	0.000000	C
10	0.000000	0.140236	0.005325	0.854439	T
11	0.000000	1.000000	0.000000	0.000000	C
12	0.000000	0.995283	0.004717	0.000000	C

Position	A	C	G	T	Motif
1	0.005096	0.712649	0.010836	0.271419	C
2	0.720253	0.008809	0.262198	0.008740	A
3	0.039639	0.005768	0.940599	0.013993	T
4	0.011739	0.917821	0.000000	0.070440	C
5	0.006753	0.915173	0.000000	0.078073	C
6	0.026904	0.014782	0.000000	0.958314	T
7	0.000000	0.923335	0.024811	0.051854	C
8	0.000013	0.938834	0.013933	0.047220	C
9	0.024899	0.799722	0.000000	0.175379	C
10	0.645698	0.013898	0.336065	0.004338	A
11	0.957413	0.005888	0.022850	0.013849	A
12	0.447738	0.000000	0.544998	0.007264	G
13	0.044533	0.016754	0.400489	0.538223	T
14	0.546965	0.000000	0.025918	0.427117	A
15	0.015499	0.000000	0.978802	0.005699	G
16	0.003894	0.929753	0.015582	0.050770	C
17	0.008058	0.004315	0.000000	0.987627	T
18	0.078901	0.003149	0.917949	0.000000	G
19	0.062283	0.003840	0.933878	0.000000	G
20	0.020689	0.000000	0.961745	0.017566	G
21	0.984796	0.000000	0.015204	0.000000	A
22	0.008549	0.299411	0.000000	0.692040	T
23	0.012159	0.034318	0.008444	0.945078	T

Position	A	C	G	T	Motif
1	0.012807	0.007857	0.975980	0.003356	G
2	0.000017	0.987575	0.000000	0.012408	C
3	0.012988	0.662972	0.011845	0.312195	C
4	0.194078	0.004687	0.790496	0.010739	G
5	0.028077	0.010382	0.956495	0.005047	G
6	0.008697	0.002055	0.989248	0.000000	G
7	0.000000	0.746058	0.005410	0.248532	C
8	0.294431	0.013047	0.676331	0.016191	G
9	0.005307	0.416010	0.005722	0.572961	T
10	0.155264	0.000000	0.837515	0.007221	G
11	0.020427	0.009314	0.964380	0.005879	G
12	0.000000	0.009739	0.000013	0.990247	T
13	0.013253	0.015870	0.963624	0.007253	G
14	0.027985	0.005510	0.966505	0.000000	G
15	0.000000	0.906610	0.006218	0.087172	C
16	0.093842	0.009889	0.258559	0.637711	T
17	0.000000	0.806979	0.166626	0.026396	C
18	0.743620	0.008202	0.243626	0.004551	A
19	0.005405	0.780326	0.005922	0.208346	C
20	0.185993	0.008545	0.799246	0.006215	G
21	0.000000	0.978220	0.000000	0.021780	C
22	0.006218	0.973098	0.002877	0.017807	C

Position	A	C	G	T	Motif
1	0.036156	0.015529	0.948315	0.000000	G
2	0.000000	0.145917	0.854083	0.000000	G
3	0.008861	0.862658	0.128481	0.000000	C
4	0.027214	0.193539	0.103785	0.675461	T
5	0.090530	0.092333	0.804090	0.013046	G
6	0.008861	0.027367	0.958663	0.005109	G
7	0.674785	0.099744	0.225471	0.000000	A
8	0.037122	0.043516	0.902308	0.017054	G
9	0.013626	0.095422	0.230692	0.660260	T
10	0.000000	0.060961	0.932467	0.006572	G
11	0.000000	0.939978	0.060022	0.000000	C
12	0.767684	0.046656	0.149667	0.035993	A
13	0.031912	0.045376	0.922713	0.000000	G
14	0.045622	0.137564	0.098419	0.718396	T
15	0.000000	0.039706	0.960294	0.000000	G
16	0.000000	0.099942	0.900058	0.000000	G

Position	A	C	G	T	Motif
1	0.019956	0.298579	0.094057	0.587408	T
2	0.026085	0.176936	0.770947	0.026031	G
3	0.000000	0.889173	0.100581	0.010246	C
4	0.734363	0.038882	0.205969	0.020786	A
5	0.002546	0.076869	0.920585	0.000000	G
6	0.000000	0.089997	0.270488	0.639516	T
7	0.151827	0.054364	0.786355	0.007453	G
8	0.732572	0.070640	0.193057	0.003731	A
9	0.000000	0.098900	0.901100	0.000000	G
10	0.000000	0.943653	0.046740	0.009607	C
11	0.074367	0.625355	0.175947	0.124331	C
12	0.012313	0.082441	0.901515	0.003731	G
13	0.770585	0.077322	0.138744	0.013349	A
14	0.000000	0.066827	0.933173	0.000000	G
15	0.622265	0.196854	0.180881	0.000000	A

A.6 Background Model RP4

Position	A	C	G	T	Motif
1	0.000828	0.002429	0.000000	0.996742	T
2	0.995576	0.000000	0.004424	0.000000	A
3	0.752685	0.000000	0.247315	0.000000	A
4	0.000000	0.015740	0.000000	0.984260	T
5	0.000000	0.982372	0.000000	0.017628	C
6	0.005474	0.965777	0.002057	0.026692	C
7	0.000000	0.954792	0.000000	0.045208	C
8	0.996629	0.003371	0.000000	0.000000	A
9	0.043484	0.000000	0.956516	0.000000	G
10	0.011313	0.985933	0.000000	0.002755	C
11	0.527679	0.005799	0.001781	0.464741	A
12	0.468610	0.498316	0.012102	0.020972	C
13	0.002816	0.457560	0.000000	0.539623	C
14	0.005371	0.021875	0.003154	0.969599	T
15	0.007817	0.311173	0.004554	0.676456	T
16	0.144888	0.010437	0.838853	0.005823	G
17	0.017943	0.005473	0.976585	0.000000	G
18	0.017494	0.000000	0.979633	0.002872	G
19	0.985612	0.000000	0.005483	0.008905	A
20	0.030264	0.000000	0.968896	0.000840	G
21	0.022613	0.002606	0.969791	0.004990	G

Position	A	C	G	T	Motif
1	0.960802	0.007024	0.030265	0.001909	A
2	0.017543	0.003670	0.877931	0.100855	G
3	0.007113	0.959854	0.010846	0.022186	C
4	0.000000	0.980605	0.007432	0.011963	C
5	0.000000	0.008166	0.004174	0.987660	T
6	0.025545	0.000000	0.974455	0.000000	G
7	0.070825	0.000000	0.929175	0.000000	G
8	0.038036	0.401267	0.555934	0.004764	G
9	0.010936	0.748175	0.007962	0.232927	C
10	0.720423	0.003915	0.271637	0.004025	A
11	0.993694	0.000000	0.003527	0.002778	A
12	0.023268	0.920100	0.011550	0.045082	C
13	0.970430	0.000000	0.025402	0.004168	A
14	0.030411	0.098860	0.387982	0.482748	T
15	0.557795	0.009943	0.432262	0.000000	A

Position	A	C	G	T	Motif
1	0.030423	0.011233	0.955102	0.003243	G
2	0.034530	0.028803	0.924780	0.011887	G
3	0.007378	0.776840	0.010119	0.205664	C
4	0.286450	0.032298	0.667924	0.013328	G
5	0.011904	0.290005	0.027426	0.670665	T
6	0.052350	0.186973	0.759078	0.001599	G
7	0.529294	0.300409	0.010812	0.159485	A
8	0.151779	0.005645	0.842577	0.000000	G
9	0.003241	0.949591	0.016759	0.030409	C
10	0.011343	0.934512	0.018687	0.035458	C
11	0.950533	0.016041	0.027983	0.005443	A
12	0.011754	0.952061	0.007924	0.028262	C
13	0.020213	0.823766	0.024859	0.131162	C
14	0.601109	0.012342	0.377544	0.009005	A
15	0.018804	0.642256	0.019936	0.319004	C
16	0.234938	0.011692	0.735469	0.017901	G
17	0.013282	0.938961	0.018420	0.029337	C
18	0.002632	0.962397	0.005414	0.029557	C
19	0.009900	0.705337	0.002444	0.282318	C
20	0.295975	0.011272	0.669207	0.023546	G
21	0.008873	0.032897	0.958221	0.000009	G
22	0.001750	0.970638	0.008283	0.019329	C

Position	A	C	G	T	Motif
1	0.677115	0.131234	0.179535	0.012116	A
2	0.039485	0.004046	0.948961	0.007509	G
3	0.006985	0.439391	0.007243	0.546380	T
4	0.129433	0.074904	0.791894	0.003769	G
5	0.989529	0.000000	0.010471	0.000000	A
6	0.003392	0.000000	0.004378	0.992230	T
7	0.000000	0.503778	0.003877	0.492345	C
8	0.002933	0.880463	0.001685	0.114920	C
9	0.155506	0.002458	0.197369	0.644667	T
10	0.000000	0.956848	0.000000	0.043152	C
11	0.018924	0.939690	0.010350	0.031036	C
12	0.008103	0.330401	0.004440	0.657056	T
13	0.177815	0.003218	0.812178	0.006789	G
14	0.000000	0.971608	0.000000	0.028392	C
15	0.000000	0.938261	0.002752	0.058987	C
16	0.000000	0.007988	0.000000	0.992012	T
17	0.007347	0.829892	0.000000	0.162762	C
18	0.680928	0.013164	0.300293	0.005614	A
19	0.020718	0.002148	0.970385	0.006748	G

Position	A	C	G	T	Motif
1	0.025361	0.153313	0.817431	0.003894	G
2	0.475092	0.185955	0.338953	0.000000	A
3	0.012003	0.097133	0.883142	0.007722	G
4	0.029908	0.196145	0.766916	0.007031	G
5	0.043853	0.263081	0.171016	0.522050	T
6	0.068529	0.155415	0.226835	0.549221	T
7	0.006335	0.070078	0.886439	0.037148	G
8	0.011107	0.740555	0.230528	0.017810	C
9	0.520921	0.136555	0.277988	0.064536	A
10	0.005366	0.103835	0.888396	0.002403	G
11	0.053359	0.153281	0.204922	0.588438	T
12	0.049875	0.095418	0.845782	0.008925	G
13	0.535672	0.162624	0.276384	0.025320	A
14	0.011117	0.178872	0.791017	0.018994	G
15	0.005398	0.654970	0.305050	0.034583	C
16	0.035354	0.612596	0.275142	0.076908	C

A.7 Background Model RP5

Position	A	C	G	T	Motif
1	0.055601	0.155916	0.782787	0.005697	G
2	0.103912	0.176250	0.661378	0.058460	G
3	0.167330	0.078553	0.680378	0.073739	G
4	0.051724	0.420761	0.486218	0.041297	G
5	0.187726	0.430834	0.184187	0.197253	C
6	0.083142	0.386706	0.386997	0.143155	G
7	0.072507	0.250546	0.437964	0.238983	G
8	0.037420	0.387271	0.557392	0.017917	G
9	0.392914	0.044820	0.301004	0.261261	A
10	0.000000	0.014891	0.959713	0.025396	G
11	0.067696	0.335044	0.493426	0.103833	G
12	0.274286	0.142959	0.467612	0.115143	G
13	0.121941	0.218586	0.534948	0.124525	C
14	0.085659	0.418857	0.382169	0.113315	C
15	0.161425	0.249052	0.402988	0.186535	G
16	0.080070	0.203254	0.657647	0.059029	G
17	0.199123	0.299415	0.368774	0.132688	G
18	0.117246	0.232955	0.531639	0.118160	G
19	0.111376	0.306282	0.524737	0.057606	C
20	0.179823	0.404250	0.337061	0.078866	C
21	0.145939	0.229589	0.563749	0.060724	G
22	0.016568	0.177489	0.799681	0.006263	G

Position	A	C	G	T	Motif
1	0.000000	0.972270	0.000006	0.027724	C
2	0.000000	0.971512	0.009030	0.019458	C
3	0.002293	0.835617	0.003131	0.158959	C
4	0.677355	0.005432	0.316211	0.001002	A
5	0.984370	0.000000	0.008223	0.007407	A
6	0.529848	0.000000	0.465285	0.004867	A
7	0.024789	0.011171	0.479947	0.484093	T
8	0.474273	0.003379	0.012305	0.510042	T
9	0.008508	0.000009	0.979140	0.012343	G
10	0.000000	0.947402	0.004695	0.047903	C
11	0.000000	0.000000	0.003366	0.996634	T
12	0.048125	0.000000	0.951875	0.000000	G
13	0.018970	0.000000	0.981030	0.000000	G
14	0.024200	0.002285	0.973515	0.000000	G
15	0.982280	0.000000	0.017720	0.000000	A
16	0.000000	0.273880	0.000000	0.726120	T
17	0.001841	0.011139	0.000000	0.987020	T
18	0.992897	0.000000	0.006276	0.000827	A
19	0.000000	0.964274	0.000000	0.035726	C
20	0.988124	0.000000	0.008024	0.003852	A
21	0.023484	0.004859	0.971657	0.000000	G

Position	A	C	G	T	Motif
1	0.002605	0.008180	0.004181	0.985034	T
2	0.023564	0.002401	0.974035	0.000000	G
3	0.074349	0.004176	0.921475	0.000000	G
4	0.040738	0.391965	0.561479	0.005817	G
5	0.010954	0.744643	0.012720	0.231683	C
6	0.720841	0.003921	0.271206	0.004032	A
7	0.993263	0.000000	0.006737	0.000000	A
8	0.023308	0.915071	0.011668	0.049953	C
9	0.975304	0.000000	0.020520	0.004175	A
10	0.030462	0.099028	0.395519	0.474992	T
11	0.568042	0.009960	0.421998	0.000000	A
12	0.033727	0.000000	0.962735	0.003537	G
13	0.016670	0.394209	0.031269	0.557853	T
14	0.266143	0.013260	0.718930	0.001667	G
15	0.986403	0.000008	0.012067	0.001523	A

Position	A	C	G	T	Motif
1	0.632687	0.163196	0.184570	0.019547	A
2	0.054894	0.000000	0.937812	0.007294	G
3	0.006701	0.402063	0.006819	0.584416	T
4	0.122443	0.055733	0.818162	0.003661	G
5	0.992067	0.004691	0.003243	0.000000	A
6	0.003302	0.000000	0.000000	0.996698	T
7	0.000000	0.551979	0.002218	0.445803	C
8	0.005987	0.872480	0.001636	0.119897	C
9	0.183624	0.002387	0.209571	0.604417	T
10	0.000000	0.941305	0.006270	0.052425	C
11	0.022681	0.946192	0.008210	0.022918	C
12	0.007872	0.384937	0.002325	0.604866	T
13	0.218279	0.000000	0.770633	0.011088	G
14	0.000000	0.977677	0.000000	0.022323	C
15	0.000000	0.924878	0.004540	0.070582	C
16	0.000000	0.000000	0.000000	1.000000	T
17	0.005410	0.810218	0.000000	0.184372	C
18	0.647291	0.009884	0.338060	0.004765	A
19	0.032350	0.000000	0.964707	0.002943	G
20	0.012953	0.953996	0.000000	0.033051	C
21	0.002873	0.963929	0.000000	0.033198	C
22	0.002151	0.000000	0.000000	0.997849	T

Position	A	C	G	T	Motif
1	0.056248	0.000000	0.921436	0.022317	G
2	0.038581	0.039886	0.882061	0.039473	G
3	0.000820	0.947179	0.022924	0.029077	C
4	0.016780	0.041811	0.010774	0.930636	T
5	0.111712	0.031982	0.842801	0.013504	G
6	0.034844	0.034553	0.899223	0.031380	G
7	0.945800	0.019007	0.026057	0.009136	A
8	0.060149	0.015258	0.910559	0.014034	G
9	0.025013	0.056499	0.051835	0.866653	T
10	0.120218	0.028903	0.783522	0.067357	G
11	0.015438	0.902570	0.024828	0.057164	C
12	0.953771	0.020292	0.014977	0.010960	A
13	0.194685	0.019836	0.769854	0.015624	G
14	0.004430	0.039486	0.000000	0.956083	T
15	0.034171	0.011702	0.954126	0.000000	G
16	0.056345	0.030962	0.886958	0.025735	G

Appendix B

Deciding the Order of the Background Model for Motif Detection

B.1 Introduction

The analysis presented in Chapter 3 naturally leads to the following question: How should one choose the order of the Markov model that represents the background for motif detection? An intuitively appealing and simple solution to this problem is as follows: Use the lowest of orders for which artificial (random) sequences generated from the corresponding Markov model *cannot* be distinguished, by any means whatsoever, from the original genomic sequences that were used to build the model.

Distinguishability can be established by only *one* methodology that is able to distinguish between the two sets of sequences. Conversely, indistinguishability needs to be established with reference to *all* conceivable methodologies that could, in principle, distinguish between the two sets of sequences. Establishing indistinguishability of two sets of sequences thus appears to be a theoretical impossibility. We circumvent this problem through the following approximate and somewhat *ad hoc* prescription: establish indistinguishability using *one* methodology that is able to distinguish between the two sets of sequences with *fair* success at low enough orders.

In the rest of this chapter, we outline one such methodology together with preliminary empirical results on how it behaves across model orders, and show how it may be harnessed to decide the order of the background model for motif detection. Our methodology, in a nutshell, is as follows: attempt to distinguish between the two sets of sequences (original vs. model-generated) using an agnostic hierarchical clustering method (such as `agnes`; to be explained shortly) coupled with a compression-based distance measure on strings (to be explained shortly). Recognizing that we have a known number of groups in the data, namely two (original vs. model-generated), we show that such clustering can be turned into a classification scheme. Further, using standard statistical measures, we assess the quality

of clustering and classification as a function of the model order. The results presented here clearly indicate that this is a promising direction that needs extensive investigation.

B.2 Materials and Methods

We downloaded 50 sets of sequences from the human genome using the `Random Sequence Grabber` each with 100 sequences of length 10000 bases. This is discussed in detail in Chapter 3, page 99. (see page 100 for details), Markov models of orders 0 through 6 were built (using `GenRGenS`) from another independent set consisting of 1000 sequences of 10000 bases each. Using `GenRGenS`, we generated 50 sets (of identical specifications as the downloaded sets) of artificial (random) sequences from each order of the Markov model. For each order of the Markov model, we thus have a total of 2500 combinations of the 50 downloaded sets and 50 artificial sets.

Hierarchical clustering was performed on each of these 2500 combinations. To avoid any bias arising from a predetermined ordering of the sequences in each set, we used a fresh random permutation of the 200 sequences (100 downloaded + 100 artificial) in each set. Clustering statistics were generated for each of these 2500 clusterings for each order. Cluster analysis was performed in the `R` statistical computing environment using the `cluster` and `fpc` packages. The distance measure used for clustering is the so-called *normalized compression distance* (NCD). We also devised a classification scheme based on clustering. The rest of this section provides details of the key ingredients of our cluster analysis.

B.2.1 Model-Generated Synthetic Sequences

For the sake of completeness, we describe here how random artificial DNA sequences can be generated using a Markov model of order k. Actual artificial sequences used in our analysis were generated using the tool `GenRGenS`.

Markov models have been discussed in detail in Chapter 3. To recapitulate, an order-0 Markov model specifies the probabilities p_A, p_C, p_G, p_T of occurrence of the four DNA letters `A, C, G, T` (see Figure B.1 for an example). To generate an artificial random sequence consisting of these letters in the given proportion, one needs a source of randomness. In the computational domain, such sources are called (pseudo)random number generators (RNG). The process of choosing the next letter in an artificial sequence randomly, with prespecified probabilities, is illustrated in Figure B.1. In essence, one needs to construct four events (corresponding to the four DNA letters) that occur with the prespecified probabilities p_A, p_C, p_G, p_T.

The same method can be extended to generate random artificial DNA sequences from a higher-order Markov model. For each k-mer, a k-th order Markov model specifies the conditional probabilities of the four DNA letters to follow that k-mer. Thus, one could start with an arbitrarily chosen k-mer, and an artificial DNA sequence can be generated by

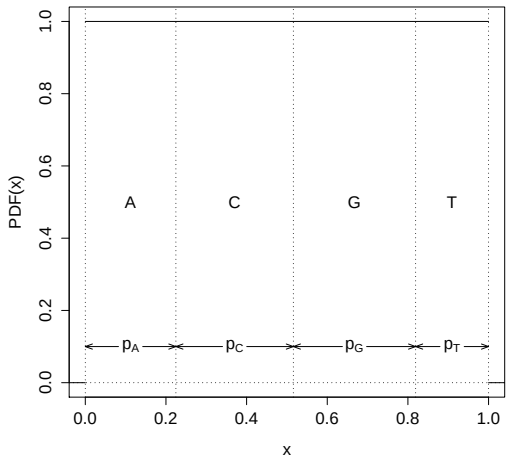


Figure B.1: **Random Generation of Artificial Sequences Using an Order-0 Markov Model**: Given a (pseudo)random number generator that generates random numbers between 0 and 1 with uniform probability density (flat solid line in the figure), a sequence of letters `A,C,G,T` with pre-specified probabilities p_A, p_C, p_G, p_T can be generated by the construction illustrated above: Generate one random number $0 \leq r < 1$; if $0 \leq r \leq p_A$, emit letter `A`; else if $p_A \leq r \leq p_A + p_C$, emit letter `C`; else if $p_A + p_C \leq r \leq p_A + p_C + p_G$, emit letter `G`; else emit letter `T`. Actual values of p_A, p_C, p_G, p_T used in the figure above are the same as those in Figure 3.1 (page 88).

generating the next letter by looking at the conditional probabilities of the four DNA letters for the last k-mer in the sequence generated, in exactly the same fashion as above.

B.2.2 Compression-Based Distance Measures for Strings

The normalized compression distance (NCD) between two strings x and y is defined as

$$\mathrm{NCD}(x, y) = \frac{C(xy) - \min(C(x), C(y))}{\max(C(x), C(y))}, \tag{B.1}$$

where $C(x)$ is the compressed size of x, xy stands for the concatenation of x followed by y, and min (max) stands for the minimum (maximum) of its arguments. The NCD has its root in the notion of Kolmogorov complexity (1) of a string, which is, qualitatively, the length of the shortest possible representation of the string. Strings with periodic, repeating patterns are thus examples of low complexity, whereas a string with no apparent pattern possessed a high complexity. While Kolmogorov complexity is uncomputable *in principle*, it is possible to *estimate* it using practical compressors such as `LZ77` or `LZ78` (2). We note in passing that several other similar distance measures, such as Chen-Li metric (CLM), compression-based dissimilarity measure (CDM), and *compression-based cosine* (CosS) have been reported in the literature (3, and references therein).

Clustering based on the NCD has been successfully used for a variety of problems such as

the classification of languages (4), authorship attribution (1), musicology (2), phylogenomics (2), classification of prokayotes (5), etc.. To cite Cilibrasi & Vitańyi (2):

> *Clustering according to NCD will group sequences together that are similar according to features that are explicitly known to us. Analysis of what the compressor actually does, still may not tell us which features that make to us can be expressed by conglomerates of features analyzed by the compressor. This can be exploited to track unknown features implicitly in classification: forming automatically clusters of data and see in which cluster (if any) a new candidate is placed.*

The versatility of the NCD perhaps originates in the fact that the algorithmic complexity theory (6) is a deep subject that may have implications *everywhere*: In essence, algorithmic complexity is closely related to the nature of correlations in a string of objects, and practical compressors attempt to recognize and represent these correlations in some (indirect) manner.

A distance measure $d(x, y)$ is expected to possess certain properties, namely,

1. $d(x, y) > 0$ for $x \neq y$ (distance between dissimilar objects is expected to be positive),
2. $d(x, x) = 0$ (distance between identical objects should be zero),
3. $d(x, y) = d(y, x)$ (symmetry), and
4. $d(x, y) \leq d(x, z) + d(z, y)$ (the triangle inequality).

Formal analysis of the NCD with reference to these properties can be found in (2). Note that with practical compressors such as `gzip` or `bzip2` (7), properties 2–4 may not always hold true. It is important that properties 2 and 3 hold if NCD is to be coupled with a hierarchical clustering algorithm. We thus use a slightly altered form of the NCD, which we call the symmetrized NCD (`SNCD`), which is manifestly symmetric and is zero for identical strings. The `SNCD` is defined as

$$\mathrm{SNCD}(x, y) = \frac{1}{2}\left[\mathrm{NCD}(x, y) + \mathrm{NCD}(y, x)\right] - \frac{1}{2}\left[\mathrm{NCD}(x, x) + \mathrm{NCD}(y, y)\right] \tag{B.2}$$

We used a home-brewn tool based on the `gzip` compression library `zlib` (7) to calculate `SNCD` for each pair of sequences in a given `FASTA` file and output in the form of a matrix (often referred to as the *dissimilarity matrix*).

B.2.3 Hierarchical Clustering

We used a hierarchical clustering method called *agglomerative nesting*. In `R`, this method is available as the function `agnes()` (8) (package `cluster`) that can take a dissimilarity matrix (in our case, the `SNCD`matrix) as input. The way `agnes` works is as follows; we quote from the `R` manual (9) page for `agnes`:

> *The `agnes`-algorithm constructs a hierarchy of clusterings. At first, each observation is a small cluster by itself. Clusters are merged until only one large cluster remains which contains*

> *all the observations. At each stage the two "nearest" clusters are combined to form one larger cluster.*

In addition, we used Ward's criterion (10) for clustering with `agnes`. To quote (10):

> *The Ward method is often successfully used for solving clustering problems over dissimilarity matrices which do not consist of squared Euclidean distances between units.*

In passing, we note here that we also attempted clustering via multiple sequence alignment using `clustalw` (11). This tool was able to distinguish biological sequences from artificial sequences generated from an order-0 Markov model. All clusterings of biological sequences with artificial sequences generated from higher-order Markov models had no clear cluster structure.

B.2.4 Statistical Measures for Quality of Clustering

Clustering results need to be assessed for their quality and stability. This is typically done by looking at the values of a variety of statistical measures. Some of the most common measures of clustering stability are described below; for a detailed explanation, see, e.g., (8).

Agglomerative coefficient (AC) is a dimensionless quantity that varies between 0 and 1. AC close to 1 indicates that a very clear cluster structure has been found in the data, whereas AC close to 0 indicates that the method has not found any natural structure in the data (in other words, the algorithm sees the data as one big cluster).

Average Silhouette Width (ASW). Values of the ASW close to +1 indicate a clear and correct clustering of the data. Values close to zero imply overlapping clusters without clear boundaries, and negative values indicate incorrect cluster assignments.

Hubert Γ. This statistic (12) measures the stability of the clustering using the instances that are clustered. Higher the value of Hubert's Γ, better the clustering.

Dunn Index (DI) (13) is the ratio of minimum separation between clusters to the maximum diameter across clusters. Large values of this index indicate better clustering.

Average Cluster Distances. Good cluster structure corresponds to a small average within-cluster distance (AWCD) and a large average between-clusters distance (ABCD). The ratio of these two quantities, called the W-B ratio (WBR), is a more useful quantity as it brings out the contrast in the cluster structure: Cleaner the clustering structure in the data, smaller the WBR.

These measures of quality and stability of clustering were computed (using the function `cluster.stats()` in the `R` package `fpc`) for each of the 2500 data combinations (see Section B.2) for each Markov model order. This enabled us to get a feel for the variability in our cluster analysis and, specifically, the dispersion of these measures of quality and stability. We estimated the dispersion of a quantity using three different estimators, namely, the standard

deviation/error (SD), the the median average deviation (MAD), and the interquantile range (IQR). The standard deviation is a standard measure of dispersion, but it is sensitive to the presence of outliers. MAD and IQR, on the other hand, are relatively robust measures of dispersion that are not affected by outliers very much. We expect that the three measures taken together give a representative picture of the true dispersion of a quantity.

In order to understand and interprete the values of these measures, we performed cluster analysis on artificial distance matrices with a known cluster structure, and tunable contrast between the within-group and between-groups distances. This contrast level was chosen to match the approximate contrast level in typical `SNCD` distance matrices for a set of 100 biological sequences and 100 artificial sequences generated from the order-0 model.

B.2.5 From Clustering to Classification

So far, we have ignored the fact that we expect two clear groups in our sequence data, namely, the group consisting of biological sequences (labelled P) picked from random locations in the human genome, and the group consisting of artificial random sequences generated from a Markov model (labelled G). Using this information is especially important because the `SNCD` distance matrices turn out to have a low contrast between the two groups of sequences. In this light, the quality of clustering could be assessed better through the following post-clustering treatment:

1. Construct two equal-sized groups of the clustered sequences. If clustering is perfect, we expect all the Ps to be in one group, and all the Gs in the other.
2. Assign to each group the label (P or G) of the sequence type that is most dominant in that group.
3. This is, in essence, a binary classification scheme. Assess the quality of such classification by calculating the Matthew correlation coefficient (MCC) (14), which is explained below.

The MCC is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{B.3}$$

where TP, TN, FP, FN respectively stand for the numbers of type-P sequences clustered with group P, type-G sequences clustered with group G, type-G sequences clustered with group P, and type-P sequences clustered with group G. In our case, by construction, we have $TP = TN$ and $FP = FN$. Equation B.3 thus reduces to the form

$$MCC = \frac{TP - FP}{TP + FP}. \tag{B.4}$$

MCC takes values between $+1$ and -1, with $+1$ implying a perfect classification, 0 implying a random assignment of group labels, and -1 meaning perfect anti-classification. We expect

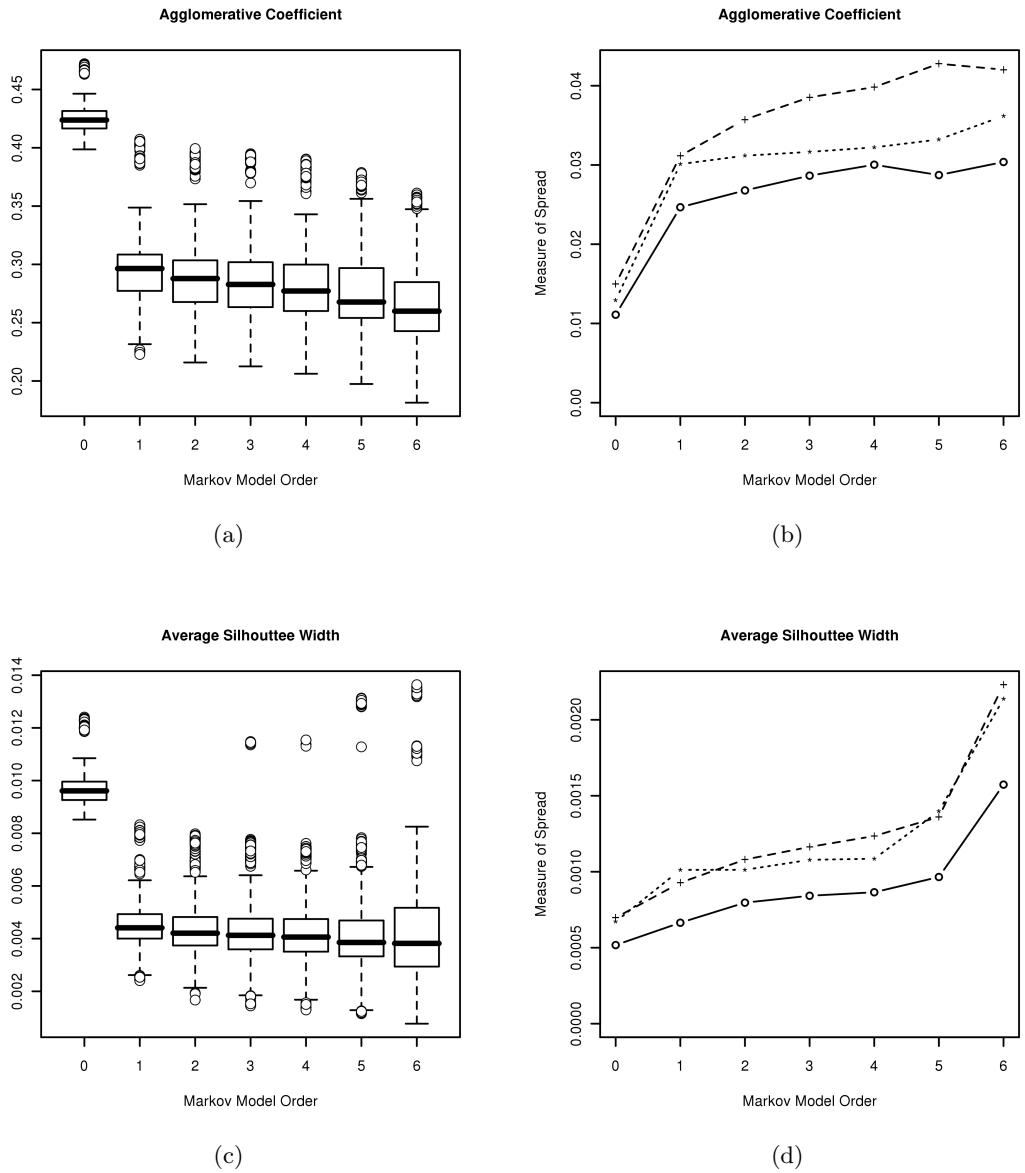


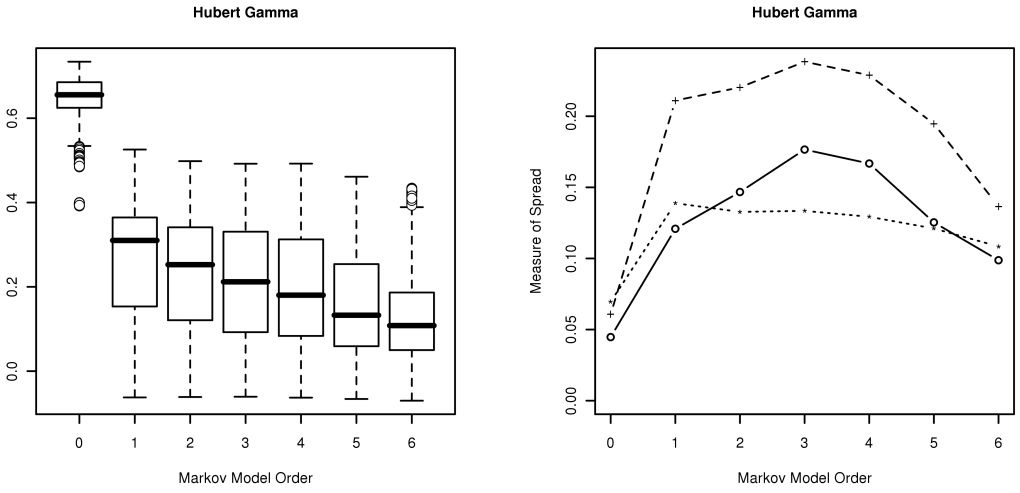
Figure B.2: Cluster Statistics I. See text for details.

the MCC to be a better measure of clustering quality in the present context, especially when the `SNCD` is able to discriminate between the genomic sequences and generated sequences only weakly.

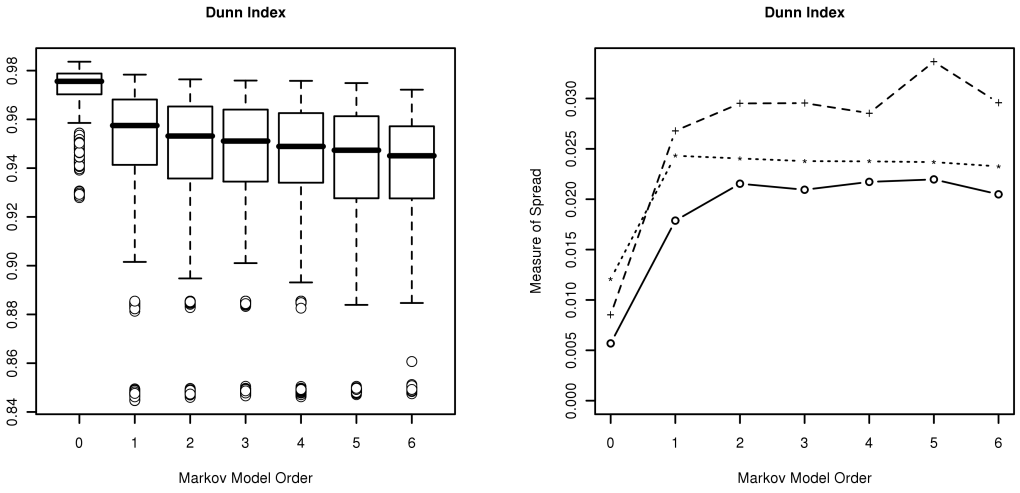
B.3 Results and Discussion

B.3.1 Clustering

Figure B.3.1 shows a representative dendrogram at order 0. We see that the P (marked blue) and the G (marked orange) sequences fall into correct groups, indicating that it is indeed possible to obtain clean and correct clustering for order 0 using the `SNCD` distance measure. In fact, such clean clustering was obtained for *all* the 2500 sequence set combinations for order 0.

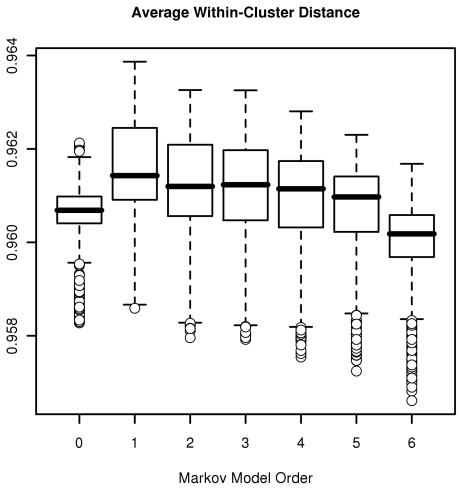


(a) (b)

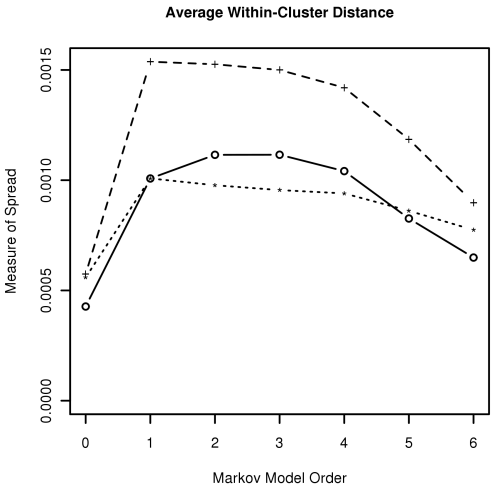


(c) (d)

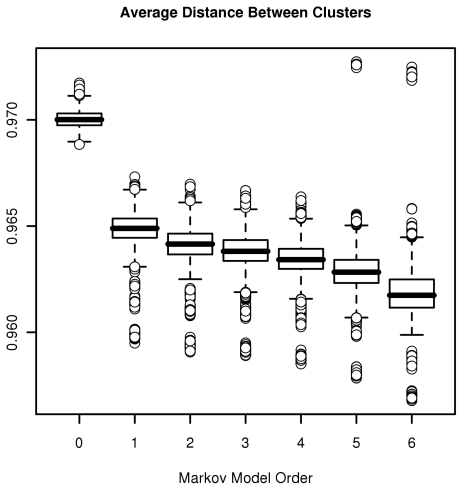
Figure B.3: Cluster Statistics II. See text for details.



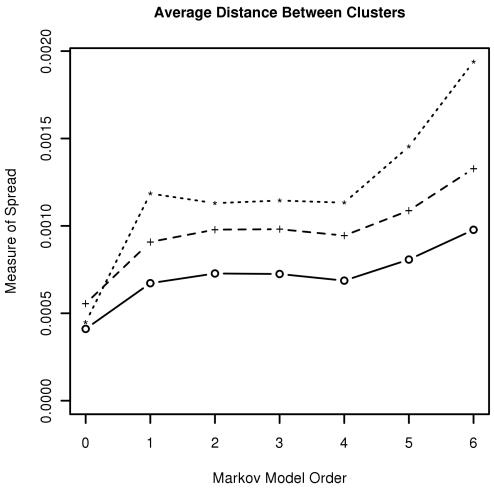
(a)



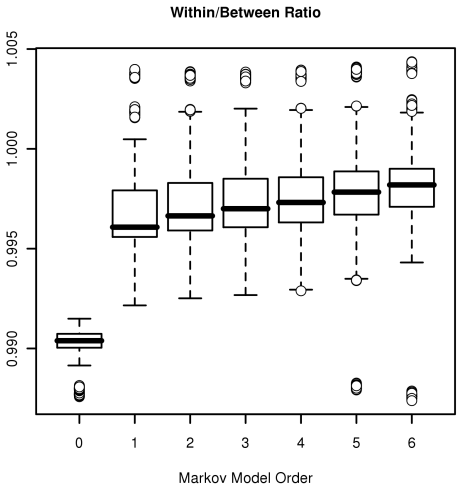
(b)



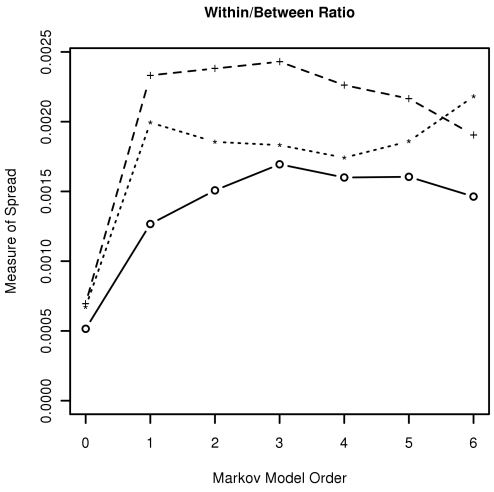
(c)



(d)



(e)



(f)

Figure B.4: Cluster Statistics III. See text for details.

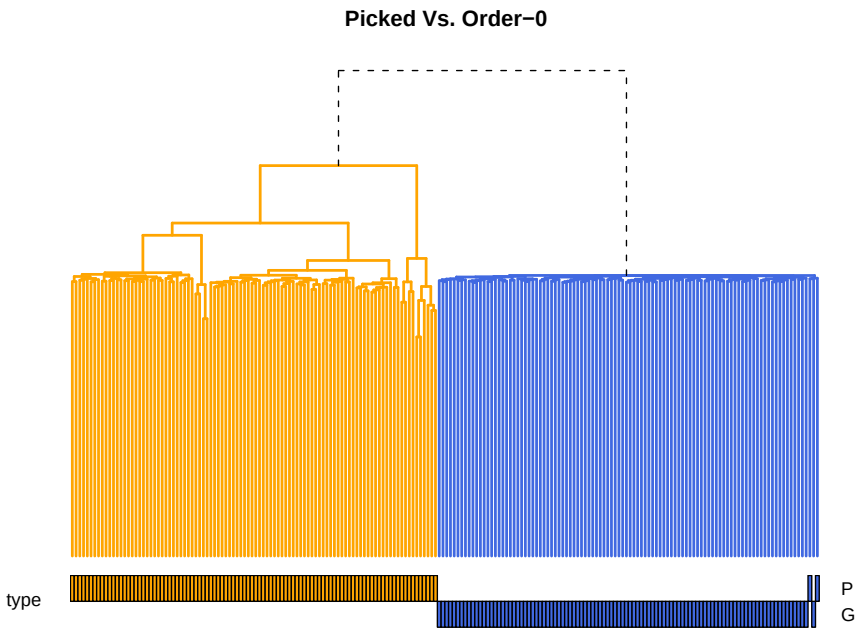


Figure B.5: **Clustering Dendrogram**: Showing the clustering of the genomic sequences and Markov order-0 generated sequences using the `agnes` with Ward. The orange labels denote the picked sequences and the blue labels show the artificial random sequences. The `agnes` with Ward is able to cluster the sequences nearly completely.

As a general rule, clustering quality deteriorates with increasing order of the Markov model.

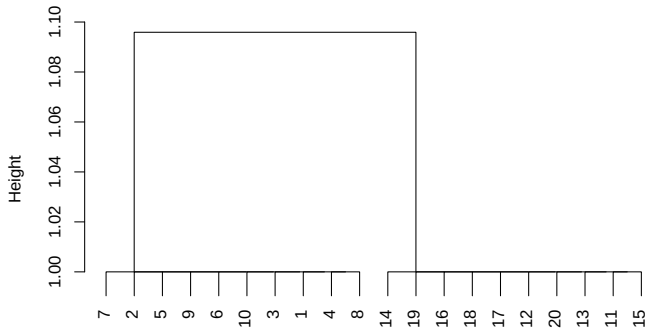
Figures B.2–B.4 show the distribution of a number of measures of clustering quality (see Section B.2.4 for details) in the form of boxplots (left-hand panels), together with their dispersions (right-hand panels) as estimated through the IQR (dashed line with a '+'), the SD (dotted line with a '$*$'), and the MAD (solid line with open circles '$\circ$'), as functions of the order of the Markov model.

The order-0 results are important for understanding our clustering results as a whole. Specifically, while we see a clean and correct clustering at order 0 in the dendrogram (Figure B.3.1), we also notice that clustering quality indicators are not in their respective ranges that are normally taken to imply a good clustering. For instance, we would have liked to see the AC to be close to 1; what we get for order 0 is rather small, around 0.4. Similarly, a value of ASW close to 1 would have been considered the indicator of good clustering; what we observe is a median value of around 0.01.

Understanding and Interpreting These Results. How does one understand and reconcile these two apparently contradictory facts, i.e., a clean and correct clustering at order 0 leading to a rather small value of the AC or the ASW? We find a clue in the values of the within-between ratio (Figure B.4(e)): values of the WBR close to 1 indicate a weak contrast of about 1% between the average within-cluster distances and average between-clusters

0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00
1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00
1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00	1.00
1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00	1.00
1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00	1.00
1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	1.00
1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00

(a) The distance matrix with known low-contrast.



(b) Dendrogram for the distance matrix in (a).

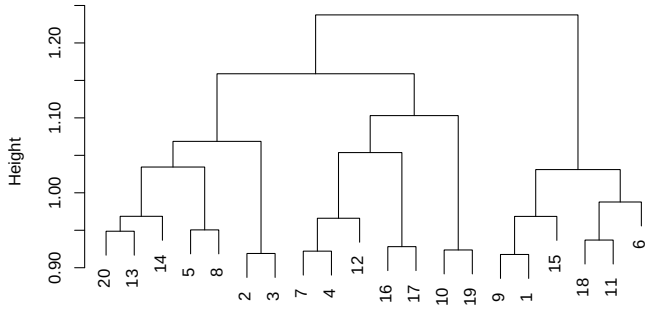
Measure	Value
AC	0.088
WBR	0.990
Hubert Γ	1.000
DI	1.010
AWS	0.010

(c) Measures of clustering quaility for clustering of distance matrix (a) and clustering in (b).

Figure B.6: Behaviour of `agnes` with low contrast data `l`. Note that the clustering is clean but the measures of quality of clustering indicate poor clustering (see text for details).

0.000	0.937	1.037	0.964	1.056	1.035	0.992	1.037	0.918	1.068	0.948	1.034	1.012	1.083	0.985	1.061	1.056	0.964	0.950	0.990
0.937	0.000	0.919	1.042	1.091	1.019	1.024	0.953	1.006	1.048	0.959	0.950	1.021	0.967	1.070	1.015	1.010	0.990	1.020	1.004
1.037	0.919	0.000	0.942	0.927	1.051	1.038	0.965	1.042	1.068	1.039	0.977	0.942	1.039	1.026	1.012	0.938	0.975	0.974	1.038
0.964	1.042	0.942	0.000	1.031	1.042	0.922	0.936	0.992	0.973	0.957	0.978	1.010	1.042	0.965	0.965	0.976	1.026	0.997	1.003
1.056	1.091	0.927	1.031	0.000	1.019	0.974	0.950	1.010	1.009	1.029	1.015	0.959	1.008	1.030	0.964	1.019	1.060	1.029	0.983
1.035	1.019	1.051	1.042	1.019	0.000	0.948	0.965	0.990	0.935	0.962	0.997	1.034	1.066	1.002	1.003	0.945	0.989	1.034	1.033
0.992	1.024	1.038	0.922	0.974	0.948	0.000	1.067	1.020	1.002	1.062	0.932	0.981	1.079	1.043	0.981	0.945	0.962	0.987	0.996
1.037	0.953	0.965	0.936	0.950	0.965	1.067	0.000	1.019	1.030	1.055	0.999	0.996	0.959	1.064	1.016	1.048	0.982	1.037	1.028
0.918	1.006	1.042	0.992	1.010	0.990	1.020	1.019	0.000	1.079	0.934	0.957	1.090	1.020	0.926	0.982	1.028	1.010	1.081	1.023
1.068	1.048	1.068	0.973	1.009	0.935	1.002	1.030	1.079	0.000	1.049	0.963	1.071	0.998	1.038	1.015	1.080	0.998	0.924	1.004
0.948	0.959	1.039	0.957	1.029	0.962	1.062	1.055	0.934	1.049	0.000	1.022	1.012	0.971	0.987	1.016	1.002	0.937	0.962	1.028
1.034	0.950	0.977	0.978	1.015	0.997	0.932	0.999	0.957	0.963	1.022	0.000	0.956	0.990	1.005	1.043	1.012	0.977	1.044	1.000
1.012	1.021	0.942	1.010	0.959	1.034	0.981	0.996	1.090	1.071	1.012	0.956	0.000	0.966	1.030	1.040	0.954	1.081	1.058	0.949
1.083	0.967	1.039	1.042	1.008	1.066	1.079	0.959	1.020	0.998	0.971	0.990	0.966	0.000	1.090	1.002	0.983	0.983	1.080	0.962
0.985	1.070	1.026	0.965	1.030	1.002	1.043	1.064	0.926	1.038	0.987	1.005	1.030	1.090	0.000	1.019	1.076	0.944	1.016	0.966
1.061	1.015	1.012	0.965	0.964	1.003	0.981	1.016	0.982	1.015	1.016	1.043	1.040	1.002	1.019	0.000	0.928	1.073	1.023	1.055
1.056	1.010	0.938	0.976	1.019	0.945	0.945	1.048	1.028	1.080	1.002	1.012	0.954	0.983	1.076	0.928	0.000	1.026	0.984	0.981
0.964	0.990	0.975	1.026	1.060	0.989	0.962	0.982	1.010	0.998	0.937	0.977	1.081	0.983	0.944	1.073	1.026	0.000	1.022	0.995
0.950	1.020	0.974	0.997	1.029	1.034	0.987	1.037	1.081	0.924	0.962	1.044	1.058	1.080	1.016	1.023	0.984	1.022	0.000	0.989
0.990	1.004	1.038	1.003	0.983	1.033	0.996	1.028	1.023	1.004	1.028	1.000	0.949	0.962	0.966	1.055	0.981	0.995	0.989	0.000

(a) The distance matrix with known low-contrast in B.6(a) but smeared with noise.



(b) Dendrogram for the distance matrix in (a).

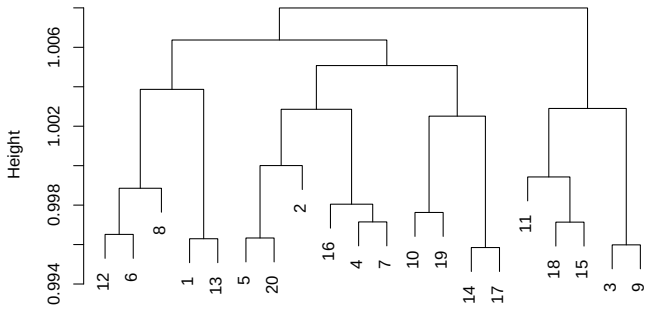
Measure	Value
AC	0.24
WBR	0.82
Hubert Γ	0.28
DI	0.86
AWS	0.03

(c) Measures of clustering quaility for clustering of distance matrix (a) and clustering in (b).

Figure B.7: Behaviour of `agnes` with low contrast data, the distance matrix is smeared with noise II. Note that the clustering is not clean, it is worst than seen in Figure B.6, but the measures of clustering quality indicate it is better than seen in Figure B.6(see text for details).

0.000	1.000	1.000	0.999	0.997	0.999	1.001	1.002	1.001	1.003	1.000	0.998	0.996	0.999	0.998	1.002	1.003	1.000	0.998	1.001
1.000	0.000	0.999	1.000	1.001	1.003	0.997	1.002	1.003	1.000	1.001	1.002	1.004	1.003	1.000	1.001	0.999	1.001	1.001	0.997
1.000	0.999	0.000	1.002	1.000	1.000	1.000	0.998	0.996	1.001	0.999	1.000	0.999	0.998	1.003	0.997	1.004	0.999	1.002	0.999
0.999	1.000	1.002	0.000	0.998	0.999	0.997	1.001	1.000	0.998	0.999	0.997	1.000	0.998	1.000	0.997	1.000	1.000	0.998	1.003
0.997	1.001	1.000	0.998	0.000	1.003	0.997	1.000	1.001	1.002	1.000	0.999	0.999	0.999	1.002	0.998	1.000	1.001	1.000	0.996
0.999	1.003	1.000	0.999	1.003	0.000	1.001	1.000	1.005	1.002	1.001	0.997	1.000	1.002	1.001	1.003	1.001	1.001	0.998	0.997
1.001	0.997	1.000	0.997	0.997	1.001	0.000	1.000	1.003	1.000	1.002	1.000	0.999	0.999	1.002	0.998	1.000	1.002	1.000	0.998
1.002	1.002	0.998	1.001	1.000	1.000	1.000	0.000	1.003	1.000	1.004	0.997	1.000	1.003	1.002	1.001	0.996	1.001	0.999	0.999
1.001	1.003	0.996	1.000	1.001	1.005	1.003	1.003	0.000	0.998	0.999	1.000	1.000	1.000	1.001	1.000	1.000	0.997	1.003	0.998
1.003	1.000	1.001	0.998	1.002	1.002	1.000	1.000	0.998	0.000	0.999	1.000	0.997	0.999	1.003	1.002	1.000	0.998	0.998	1.000
1.000	1.001	0.999	0.999	1.000	1.001	1.002	1.004	0.999	0.999	0.000	0.999	1.001	0.998	0.998	1.000	1.002	1.000	1.002	1.002
0.998	1.002	1.000	0.997	0.999	0.997	1.000	0.997	1.000	1.000	0.999	0.000	1.000	1.002	0.998	1.002	1.001	1.000	1.000	0.999
0.996	1.004	0.999	1.000	0.999	1.000	0.999	1.000	1.000	0.997	1.001	1.000	0.000	1.002	1.001	0.998	1.001	1.001	1.002	1.001
0.999	1.003	0.998	0.998	0.999	1.002	0.999	1.003	1.000	0.999	0.998	1.002	1.002	0.000	1.001	1.001	0.996	0.999	1.001	1.000
0.998	1.000	1.003	1.000	1.002	1.001	1.002	1.002	1.001	1.003	0.998	0.998	1.001	1.001	0.000	0.999	1.001	0.997	1.003	0.999
1.002	1.001	0.997	0.997	0.998	1.003	0.998	1.001	1.000	1.002	1.000	1.002	0.998	1.001	0.999	0.000	0.999	1.004	1.001	1.004
1.003	0.999	1.004	1.000	1.000	1.001	1.000	0.996	1.000	1.000	1.002	1.001	1.001	0.996	1.001	0.999	0.000	0.998	0.999	1.001
1.000	1.001	0.999	1.000	1.001	1.001	1.002	1.001	0.997	0.998	1.000	1.000	1.001	0.999	0.997	1.004	0.998	0.000	1.002	1.000
0.998	1.001	1.002	0.998	1.000	0.998	1.000	0.999	1.003	0.998	1.002	1.000	1.002	1.001	1.003	1.001	0.999	1.002	0.000	1.002
1.001	0.997	0.999	1.003	0.996	0.997	0.998	0.999	0.998	1.000	1.002	0.999	1.001	1.000	0.999	1.004	1.001	1.000	1.002	0.000

(a) Distance matrix generated using Uniform(0.995,1.005).



(b) Dendrogram for the distance matrix in (a).

Measure	Value
AC	0.01
WBR	0.99
Hubert Γ	0.21
DI	0.99
AWS	0.0009

(c) Measures of clustering quaility for clustering of distance matrix (a) and clustering in (b).

Figure B.8: Behaviour of `agnes` with low contrast data III. We expect no clustering in this case as all the sequences are generated randomly. (see text for details).

distances.

What appears to be happening is that the `SNCD` is indeed able to distinguish between the P sequences and the G sequences at order 0 (this would explain the clean clustering as seen in the dendrogram B.3.1), but with a rather weak contrast (this would explain why clustering quality indicators are not so clean). To verify this, we performed clustering exercises on the following artificial dissimilarity matrices:

1. Dissimilarity matrices with a known small contrast between two groups of equal size (100 each). An example of such artificial dissimilarity matrix (of a smaller size) is shown in Figure B.6. Here, all within-group distances are set equal to 1, whereas all distances between members of one group to members of the other are set to 1.01. This contrast level of 1% is chosen to match the approximate contrast level in the `SNCD` dissimilarity matrices at order 0. This contrast level corresponds to a WBR of $1/1.01 \approx 0.990099$.

   The results of this exercise with smaller artificial dissimilarity matrices are shown in Figure B.6. We indeed see that clean and correct clustering can be obtained (using `agnes` with Ward) even in a low-contrast situation. The corresponding quality indicators indeed indicate a poor clustering because of the low inter-group contrast.

   Specifically, when the same exercise is done for two groups of size 100 each (dendrogram not shown), the AC turns out to be approximately 0.42, which is close to the median AC value of 0.43 that we see in our sequence clustering exercise at order 0.

2. The same dissimilarity matrix "smeared" with multiplicative random Uniform(0.9,1.1)[1] noise. This exercise is expected to give us a feel for how the clustering quality deteriorates with noise.

   We see in Figure B.7 that the underlying low-contrast cluster structure can still be discerned in presence of noise at this level. Although some data instances do get wrongly assigned to the opposite group, most data, by and large, gets correctly clustered. The value of AC, although smaller than in the unsmeared case, is still comparable.

   We see the same behavior/trend for two groups of size 100 each (AC ≈ 0.38).

3. For the sake of completeness, we also performed clustering with an artificial dissimilarity matrix that consisted of Uniform(0.995,1.005) noise alone; we expect to see only one group in the data, and the value of AC is very close to 0, as expected (See Figure B.8).

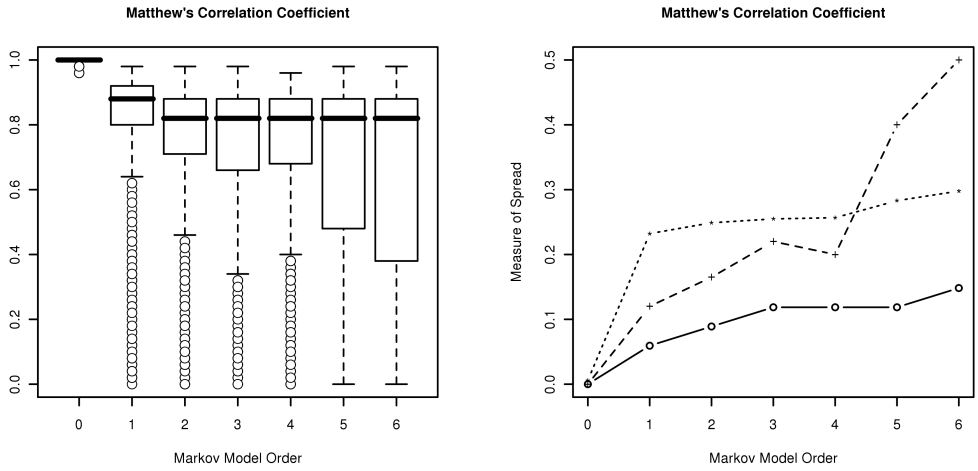
   We again see the same behavior/trend for two groups of size 100 each (AC ≈ 0.03).

We thus conclude that the quality of `agnes`/Ward clustering in a low-contrast situation may very well be good even if the standard measures of clustering quality indicate otherwise.

B.3.2 Classification

In arriving at the above conclusion, we have made use of additional information that we expect two equal-sized groups in the sequence data. With this additional information, the quality

[1] Uniform(a,b) stands for the uniform distribution over the interval a to b.



(a) **Matthews Correlation Coefficient** See text for details.

(b) The spread in the values of MCC in terms of measures of statistical dispersion.

Figure B.9: The panel on the left-hand side denotes boxplot for a given cluster statistic as plotted over 2500 clusterings, each panel on the right-hand side denotes the spread for the respective boxplots in terms of 3 measures of statistical dispersion, dashed line with '+' mark denotes the IQR (interquantile range), the dotted line with '*' denotes the standard deviation and the line with open circles '∘' denotes the MAD (median average deviation).

of such binary clustering can be assessed in a better fashion, especially in a low-contrast situation, via the Matthew correlation coefficient defined in Section B.2.5. The behavior of the MCC as a function of the order of the Markov model is displayed in Figure B.9. We see that the value of the MCC for order 0 reflects the near-perfect classification into P and G groups. As expected, classification accuracy deteriorates at higher orders as model-generated sequences start resembling genomic sequences better. Interestingly, while the median value of the MCC (see Figure B.9(a)) does not decrease drastically with order, the dispersion (see Figure B.9(b)) of the MCC increases dramatically.

This behavior suggests that the order of the background model in a motif detection exercise can be decided on the basis of the dispersion of the MCC. Our prescription for the background model order is thus as follows: choose the lowest order with dispersion of MCC larger than a prespecified threshold.

B.3.3 Summary and Conclusions

The focus of this work was to arrive at a prescription for deciding the order the background model for motif detection. Given that an appropriate background has been identified in the context of the motif detection problem, an intuitively appealing and simple prescription is as follows: Use the lowest of orders for which artificial (random) sequences generated from the corresponding Markov model *cannot* be distinguished, by any means whatsoever, from the background genomic sequences that were used to build the model. To establish

indistinguishability, we suggested the use of `SNCD`-based agnostic clustering of background genomic sequences and artificial model-generated sequences. Through an extensive clustering exercise, we established that the `SNCD` distance measure coupled with `agnes`+Ward clustering is able to distinguish systematically between the two sets of sequences unambiguously at order 0, even if the contrast between within-group and between-groups distances is rather low, of the order of 1%. Based on an analysis of these results, we devised a classification scheme that captures the quality of clustering correctly through the Matthew correlation coefficient.

Our complete prescription for determining the order of the background model for motif detection is as follows:

1. Identify the correct background based on the problem at hand. Collect $M + 1$ sets of N of (genomic) sequences of length L each from this background.
2. Build Markov models of orders $0, 1, 2, \ldots$ using *one* of these sets of sequences.
3. For each such model, generate M sets of N artificial random sequences of length L each.
4. For each of the M^2 combinations (of M remaining genomic sets with M artificial sets), compute the `SNCD` dissimilarity matrix. Cluster each combination using `agnes` with Ward.
5. Calculate the MCC for each of the resulting clusterings (see Section B.2.5), and compute some measure of the spread of its distribution.
6. Using some reasonable threshold on the spread of MCC, choose the lowest order for which the spread of MCC is above this threshold. This is the prescribed order of the Markov model representing the background in a motif detection exercise.

We note in passing that any better distance measure could be substituted for `SNCD`, and any better clustering method could be substituted for `agnes` with Ward; however, the behavior of such a combination needs to be thoroughly investigated prior to use.

References

[1] Emanuele Caglioti Andrea Baronchelli and Vittorio Loreto. Artificial sequences and complexity measures. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(04):P04002, 2005.

[2] Rudi Cilibrasi and Paul M. B. Vitańyi. Clustering by Compression. *IEEE Trans. Inform. Th.*, 51(4):1523–1545, April 2005.

[3] Sculley D and Carla E. Brodley. Compression and Machine Learning:A New Perspective on Feature Space Vectors. *Data Compression Conference*, 2006.

[4] D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. *Phys Rev Lett*, 88(4):048702, 2002.

[5] P. Ferragina, R. Giancarlo, V. Greco, G. Manzini, and G. Valiente. Compression-based classification of biological sequences and structures via the Universal Similarity Metric: experimental assessment. *BMC Bioinformatics*, 8:252, 2007.

[6] Gregory J. Chaitin. *Thinking about Gödel and Turing: Essays on Complexity, 1970-2007*. World Scientific, 2007.

[7] Jacob Ziv and Abraham Lempel. Compression of Individual Sequences via Variable-Rate Coding. *IEEE TRANSACTIONS ON INFORMATION THEORY*, IT-24(5):530–536, 1978.

[8] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data:An Introduction to Cluster Analysis*. Wiley, New York, 1990.

[9] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0.

[10] M. R. Anderberg. *Cluster Analysis for Applications*. Academic Press, New York, 1973.

[11] W. R. Pearson. Using the FASTA program to search protein and DNA sequence databases. *Methods Mol Biol*, 24:307–31, 1994.

[12] Hubert Lawrence. Monotone invariant clustering procedures. *Psychometrika*, 38:47 – 62, 1973.

[13] M. Haldiki, Y. Batistakis, and M. Vazirgiannis. Cluster validity methods, Part I. *SIGMOD Record*, 31:40 – 45, 2002.

[14] B. W. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, 405(2):442–51, 1975.

Appendix C

Other Data

This part of the thesis is provided as a DVD inside the last page cover fold of the thesis. On that DVD there is a `README` file that describes the data available in the DVD. The folders are arranged chapter-wise (as they appear in the thesis) and all the relevant information is clubbed together in various folders.
If there are problems in reading the DVD or accessing the DVD, all the data in the thesis is available on request. Kindly direct your request to `sameet@cms.unipune.ernet.in`.